Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Dynamic Batch Learning in High-Dimensional Sparse Linear Contextual Bandits

Zhimei Ren, Zhengyuan Zhou

To cite this article:
Zhimei Ren, Zhengyuan Zhou (2024) Dynamic Batch Learning in High-Dimensional Sparse Linear Contextual Bandits. Management Science 70(2):1315-1342. https://doi.org/10.1287/mnsc.2023.4895

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 70, No. 2, February 2024, pp. 1315–1342
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Dynamic Batch Learning in High-Dimensional Sparse Linear Contextual Bandits

Zhimei Ren,[a] Zhengyuan Zhou[b,*]

[a] Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [b] Department of Technology, Operation and Statistics, Stern School of Business, New York University, New York, New York 10012

*Corresponding author

Contact: zren@wharton.upenn.edu, https://orcid.org/0000-0002-2872-5842 (ZR); zzhou@stern.nyu.edu, https://orcid.org/0000-0002-0005-9411 (ZZ)

Received: May 10, 2021
Revised: June 22, 2022
Accepted: October 10, 2022
Published Online in Articles in Advance: October 19, 2023

https://doi.org/10.1287/mnsc.2023.4895

Abstract. We study the problem of dynamic batch learning in high-dimensional sparse linear contextual bandits, where a decision maker, under a given maximum-number-of-batch constraint and only able to observe rewards at the end of each batch, can dynamically decide how many individuals to include in the next batch (at the end of the current batch) and what personalized action-selection scheme to adopt within each batch. Such batch constraints are ubiquitous in a variety of practical contexts, including personalized product offerings in marketing and medical treatment selection in clinical trials. We characterize the fundamental learning limit in this problem via a regret lower bound and provide a matching upper bound (up to log factors), thus prescribing an optimal scheme for this problem. To the best of our knowledge, our work provides the first inroad into a theoretical understanding of dynamic batch learning in high-dimensional sparse linear contextual bandits. Notably, even a special case of our result—when no batch constraint is present—yields that the simple exploration-free algorithm using the LASSO estimator already achieves the minimax optimal $\tilde{O}(\sqrt{s_0 T})$ regret bound (s_0 is the sparsity parameter or an upper bound thereof and T is the learning horizon) for standard online learning in high-dimensional linear contextual bandits (for the no-margin case), a result that appears unknown in the emerging literature of high-dimensional contextual bandits.

History: Accepted by Baris Ata, stochastic models and simulation.

Funding: This work is supported by the National Science Foundation [Grant CCF-2106508]. Z. Zhou gratefully acknowledges the Digital Twin research grant from Bain & Company and the New York University's 2022-2023 Center for Global Economy and Business faculty research grant for support on this work. Z. Ren was supported by the National Science Foundation [Grant OAC 1934578] and by the Discovery Innovation Fund for Biomedical Data Sciences.

Keywords: dynamic batch learning • LASSO • high-dimensional statistics • contextual bandits • sparsity

1. Introduction

With the growing abundance of user-specific data, service personalization—tailoring service decisions based on each individual's own characteristics—has emerged to be a predominant paradigm in data-driven decision making. This is because through personalization, a decision maker can exploit the heterogeneity in a given population by selecting the best decisions on a fine-grained individual level, thereby improving the outcomes. Such heterogeneity is ubiquitous; and intelligently capturing its benefits through personalization has found immense benefits across a wide range of applications in operations management, including medical treatment selection in clinical trials, product recommendation in marketing, ads selection in online advertising and nurse staffing in hospital operating rooms (Bertsimas and Mersereau 2007, Kim et al. 2011, Bastani et al. 2017, Mintz et al. 2017, Schwartz et al. 2017, Ferreira et al. 2018, Hopp et al. 2018, Zhou et al. 2018, Ban and Rudin 2019, Miao and Chao 2019).

In the current era, such data-driven personalized decision-making problems often exhibit both high-dimensionality and sparsity (Naik et al. 2008, Belloni and Chernozhukov 2011, Kim et al. 2011, Bayati et al. 2014, Belloni et al. 2014, Razavian et al. 2015, Zhou et al. 2018). High-dimensionality refers to the fact that, as a result of modern data collection technologies, a large number of features about individuals are collected and recorded in the data sets, hence making the covariate vector high-dimensional. At the same time, the underlying reward response model is often *sparse*, where only a few of those covariates actually influence the rewards. To capture these two aspects, and to take into account the sequential decision making nature of personalization, such problems have been formalized in the framework of high-dimensional sparse linear

contextual bandits, where the contexts are independently and identically distributed (i.i.d.) drawn from an underlying distribution and the context dimension d is comparable or even exceeds the learning horizon T, whereas at most s_0 ($\ll d$) context variables influence the (random) reward, which in expectation is a linear function of the context vector.

Driven by a pressing need to achieve effective personalization in this challenging regime, an emerging line of work (Wang et al. 2018, Kim and Paik 2019, Bastani and Bayati 2020) has developed algorithms and established regret guarantees, where regret measures the performance difference between the cumulative reward generated by the algorithm and that achieved by an optimal policy (if the underlying model were known). This line of work has exploited the fact that the underlying linear model is sparse to achieve regret bounds that scale gracefully with s_0 (much smaller than the ambient dimension d). For instance, under further margin conditions (where a gap between the optimal action and suboptimal actions can be identified with positive probability and where regret logarithmic in T is thus possible), Bastani and Bayati (2020) have developed a forced-sampling exploration scheme that is used jointly with the least absolute shrinkage and selection operator (LASSO) estimator and established a $O(s_0^2 \cdot (\log d + \log T)^2)$ regret bound. Building on Bastani and Bayati (2020), Wang et al. (2018) then subsequently[1] used the same forced-sampling exploration scheme, but with a different minimax concave penalty weighted LASSO estimator and obtained the $O(s_0^2 \cdot (\log d + s_0) \cdot \log T)$ regret bound, an improvement if s_0 is not much larger compared with $\log T$ and/or $\log d$. When no margin condition exists (in which case dependence on T is at best $\Omega(\sqrt{T})$), Kim and Paik (2019) constructed a doubly robust LASSO estimator based algorithm (with uniform sampling exploration) that achieves $\tilde{O}(s_0\sqrt{T})$ regret. Additionally, several earlier works (Abbasi-Yadkori 2012, Carpentier and Munos 2012) also studied high-dimensional linear contextual bandits but did not use LASSO based methods: These methods are either restricted to specialized settings—special action set structure and nonstandard noise in Carpentier and Munos (2012)—or obtained regret bounds that are worse[2] than $\tilde{\Theta}(\sqrt{dT})$. Taken together, these developments represent fruitful inroads into the high-dimensional regimes that intelligently exploited sparsity for practical benefits.

Despite these fruitful studies, an important aspect is missing in this line of work that limits their applicability in practice. The standard online learning model adopted in the literature—where a decision is made on the current individual, yielding an outcome that is immediately observed and incorporated to make the next decision—is simply impractical in many applications. In practice, although decision makers are able to perform active learning and incorporate feedback from the past to adapt their decisions in the future, such adaptation is often limited to a fixed number of rounds of interaction due to physical, cost, or regulatory constraints. We refer to this constraint as the batch constraint in this paper. For instance, when running a personalized product marketing campaign—a prime example where high-dimensional customer data are available (Bertsimas and Mersereau 2007, Schwartz et al. 2017)—a company often needs to mail personalized product offers to its (existing and/or potential) customers. Here, the marketer will not (and cannot afford to) make a product offer to one customer, wait to receive feedback and then move on to the next customer (the standard online learning model). Instead, the marketer in practice will batch mail a set of customers, receive their feedback collectively and then design the next batch of offerings accordingly. The marketer typically has a targeted customer population at hand (selected from the entire customer base) and, working with a time and monetary budget that dictates the maximum number of batches, needs to design how to optimally partition the customer population into different batches and what product to offer to each customer in a given batch.

Another example where such a batch constraint exists is adaptive clinical trials (Robbins 1952, Chow and Chang 2008, Pallmann et al. 2018), where a fixed number of medical treatments (e.g. different drugs or same drug but different dosages or both) are applied to a group of patients based on the patients' medical characteristics during a phase of the trial, with the medical outcomes collected for the entire group at the end of the phase. The data collected from previous phases are then analyzed to design the next phase, including how many patients to include for the next phase and the medical treatment assignment to the patients. Here, each phase corresponds to a batch of participating patients. As pointed out in Pallmann et al. (2018, pp. 1–15), "adaptive designs can make clinical trials more flexible by utilizing results accumulating in the trial to modify the trial's course in accordance with pre-specified rules." Despite being offered such flexibility in adaptive clinical trials (compared with traditional nonadaptive trials), medical decision makers have limited adaptivity here because the medical outcome for a patient can only be observed after a sufficient amount of time has passed; as such, the trials must proceed in batches (phases). The current U.S. Food and Drug Administration (FDA) regulation requires four phases for a standard clinical trial. Thus, the trial patients need to be partitioned into four batches, and incorporation of new information only occurs at the end of each batch, thereby rendering the standard online learning model and hence the standard online bandits/contextual bandits algorithms inapplicable. We do point out that pharmaceutical companies that conduct clinical trials often have the sole objective of obtaining FDA approval and hence do not have the objective of maximizing the total welfare (measured by regret) of the trial patients. In contrast, our paper's focus is on maximizing total welfare

(equivalently minimizing regret),and hence would shed light for adaptive clinical trials under this criterion.

From the previous statements, we see that the key challenge imposed by the batch constraint is the limited adaptability: The adaptation can only occur at a batch level rather than at an individual level. Such limited adaptability forces the decision maker to carefully select the batches, based on available information from the past, so that the inability to adapt (and hence the inferior performance resulted therefrom) does not cause much degradation to the overall performance. Motivated by these considerations, we study the problem of dynamic batch learning, where a decision maker dynamically decides the next batch's size (at the end of the current batch) and what personalization scheme to adopt within each batch under a given maximum-number-of-batch constraint.

1.1. Main Results

Our main contributions are twofold. First, we study the fundamental limits of dynamic batch learning in high-dimensional sparse linear contextual bandits. By an information-theoretical argument that carefully selects a sequence of Bayesian priors, we establish an $\Omega\Big(\max\{M^{-4}2^{-\frac{7}{2}M}\sqrt{Ts_0}(T/s_0)^{\frac{1}{2(2^M-1)}}, \sqrt{Ts_0}\}\Big)$ regret lower bound (Theorem 1), where M is the maximum number of batches allowed. This lower bound—which holds even for the simple standard Guassian contexts—indicates that regardless of how one dynamically makes partitions and/or performs action selection within each batch, the regret can never be made any smaller. For instance, if $M=4$ (as is the case for clinical trials), then no scheme can achieve better regret than $\Omega(T^{\frac{8}{15}}s_0^{\frac{7}{15}})$. The second term $\Omega(\sqrt{Ts_0})$ in the max is a lower bound[3] for the standard online learning setting, which is automatically a lower bound for dynamic batch learning since the presence of a batch constraint only makes the problem harder. Furthermore, the break-even point (up to log factors) between these two terms is $M=\Theta(\log\log(T/s_0))$, suggesting that—if the lower bound is tight—only $\Theta(\log\log(T/s_0))$ (practically a constant number) batches are needed to achieve the optimal performance of standard online learning, where no batch constraint exists.

Second, we establish that this lower bound is indeed tight (up to log factors) by providing a matching upper bound. In particular, through a simple LASSO batch greedy learning algorithm (Algorithm 1), we establish in Theorem F.2 and Theorem 2 that the regret is upper bounded by $\tilde{O}(\sqrt{Ts_0}(T/s_0)^{\frac{1}{2(2^M-1)}})$ when the number of batches does not exceed $O(\text{loglog}(T/s_0))$, hence validating that only $\Theta(\text{loglog}(T/s_0))$ batches are needed to achieve $\tilde{O}(\sqrt{Ts_0})$ regret. It suffices to look at M that is $O(\text{loglog}(T/s_0))$, because the regret will not get worse and hence will stay at $\tilde{O}(\sqrt{Ts_0})$ when M gets larger. In particular, a special case of this result (Corollary 1) is that in the standard online learning setting where no-margin exists, we can achieve the minimax optimal regret $\tilde{\Theta}(\sqrt{Ts_0})$ using an exploration-free and computationally efficient algorithm, improving on the $\tilde{O}(s_0\sqrt{T})$ regret bound given in Kim and Paik (2019).

Notably, the algorithm that achieves such strong guarantees is simple: it uses a static grid and is exploration-free. By the lower bound, using a static grid is not a limitation of the algorithm, but an attestation to its strength (easy implementability in practice). That exploration-free suffices is yet another important message, both for dynamic batch learning and standard online learning. For the latter, the existing state-of-the-art algorithms (Kim and Paik 2019, Bastani and Bayati 2020) all use contrived forced-sampling exploration schemes, which is burdensome to implement in practice. However, our results show that forced-sampling exploration is not necessary, thus echoing in high dimensions a similar message advocated in Bastani et al. (2021) for low-dimensional linear contextual bandits.

1.2. Managerial Insights

Our results provide insights on how to prescribe the *optimal* personalization scheme when limited adaptivity is present in practice and the resulting performance gap (or the absence thereof) when compared with the ideal fully online setting. These insights can help managerial decision makers in different ways, depending on the context. First, when the adaptivity constraint M is fixed a priori, such as in the clinical trials setting with $M=4$, our work provides prescriptive solutions for how to design the trials to achieve optimal performance. Furthermore, this optimal performance is $\Omega(T^{\frac{8}{15}}s_0^{\frac{7}{15}})$, whereas the infeasible fully online optimal performance is $\sqrt{Ts_0}$, a quantity that is quite close.

Second, when the limited adaptivity constraint M is not as rigid and can hence be thought as a variable subject to certain budget limit, our results contribute meaningfully to the larger cost/benefit discussions facing the managerial decision makers. For instance, in the personalized product recommendation application, our results indicate that $\text{loglog}(T/s_0)$ rounds of campaigns are needed to achieve the (practically infeasible) fully online performance (where T here corresponds to the number of customers). This is usually a very small number and the result makes it clear that the decision maker should never need to budget for more than that. On the other hand, if under a tight budget constraint (and hence unable to finance $\text{loglog}(T/s_0)$ rounds), the decision maker would be clearly informed by the particular benefits under a range of feasible M's and how to execute it optimally once such an M is decided on. Taken together, we believe our results provide valuable prescriptive insights in the area of adaptive personalization when limited adaptivity is present.

1.3. Other Related Work

The bandits literature is extensive and much of the existing work in that space study low-dimensional contextual

bandits (see Bubeck et al. (2012), Lattimore and Szepesvári (2020), and Slivkins et al. (2019) for three books on this research area), where the dimension d of the contexts is small compared with the learning horizon T and where many well-performing algorithms have been developed and strong theoretical guarantees have been established (see Filippi et al. (2010); Rigollet and Zeevi (2010); Chu et al. (2011); Goldenshluger and Zeevi (2013); Agrawal and Goyal (2013a, b); Russo and Van Roy (2016); and Mintz et al. (2020) for a highly incomplete list). Low-dimensional contextual bandits are not our focus here, and we simply mention in passing that applying (state-of-the-art) results from the low-dimensional contextual bandits literature to the high-dimensional setting often yields results that are not useful. For instance, in linear contextual bandits with no-margin, one obtains the $\tilde{\Theta}(\sqrt{dT})$ regret by applying the result in Chu et al. (2011). Even if such regret bounds continue to hold in high dimensions,[4] such performance guarantees are not meaningful anymore, because when $d = \Omega(T)$ (d could also be a lot larger than T), at least linear regret $\tilde{\Omega}(T)$ is incurred, thus yielding completely ineffective learning.

Additionally, we point out that batch-constrained learning in bandits has been studied before in the literature. In two-armed multiarmed bandits (MABs), Perchet et al. (2016) studied static batch learning where the batch sizes must be decided a prior and established that $O(\log\log T)$ batches are needed (via a successive elimination algorithm during each batch) to achieve the same regret bound as in standard online learning. Gao et al. (2019) then generalized the result to K-armed bandits (using the same algorithm) and obtained a tight $\Theta(\log\log T)$ regret bound even when the batch sizes can be chosen dynamically. However, since MABs do not capture individuals' characteristics, these initial efforts (Perchet et al. 2016, Gao et al. 2019) only operate on a population level and do not address the problem of personalized decision making, which severely limits their practical applicability. More recently, Han et al. (2020) has studied this problem in low-dimensional linear contextual bandits, and provides the first characterization of batch learning that incorporates personalized decision making. In particular, a greedy ordinary least squares based algorithm is shown to achieve optimal regret (up to log factors). Despite these strong guarantees, the results in Han et al. (2020) are insufficient for several reasons. First, importantly, the setting in Han et al. (2020) is limited to the low-dimensional regime where $d = O(\sqrt{T})$. Second, Han et al. (2020) studied static batch learning where the batch partitions must be chosen prior to the start of the decision-making process and cannot be changed thereafter. Consequently, this raises the critical issue of whether one can do better if *dynamic batch learning* (where the decision maker can decide the next partition based on the data observed thus far) is allowed, a question whose answer is not at all obvious. Third, Han et al. (2020) works exclusively with Guassian contexts, and its proofs rely on such Guassianity, which thus limits its applicability. In contrast, our goal in this paper is to delineate—in the high-dimensional sparse setting—the performance of dynamic batch learning by providing theoretical characterizations. Additionally, when restricted to the low-dimensional setting (by taking $d = s_0$) with batch constraints, our results provide a strict generalization of Han et al. (2020) on several fronts when the underlying contexts are stochastically generated (Han et al. (2020) also investigated adversarially generated contexts, which we do not study here): We study dynamic batch learning and we deal with general sub-Gaussian contexts (with diversity condition). Consequently, although our goal lies in understanding dynamic batch learning under high-dimensional sparsity, our results are also state-of-the-art in low dimensions as well.

2. Problem Formulation

We start with some useful notation that will be used throughout the paper. For a positive integer n, $[n]$ denotes the set $\{1, 2, \ldots, n\}$; $\mathbb{S}^{n-1}$ denotes the $(n-1)$-dimensional unit sphere; $\Delta\mathbb{S}^{n-1}$ denotes the $(n-1)$-dimensional sphere with radius Δ, for a given $\Delta > 0$; and $|S|$ denotes the cardinality of the set S and S^c denotes the complement of S. For a vector v and a nonnegative integer q, $\|v\|_q$ denotes the ℓ_q norm of v. For any positive semidefinite matrix A, $\lambda_{\min}(A)$ denotes its smallest eigenvalue, and $\lambda_{\max}(A)$ its largest eigenvalue. We now move on to the formulation of the problem.

2.1. High-Dimensional Sparse Linear Contextual Bandits

Let T denote the time horizon, d the feature dimension and K the number of arms. At $t \in [T]$, the decision maker first observes a set of K d-dimensional feature vectors (i.e., contexts) $\{x_{t,a}\}_{a\in[K]}$. If the decision maker selects action $a \in [K]$, then a reward $r_{t,a} \in \mathbb{R}$ is incurred: $r_{t,a} = x_{t,a}^\top \theta^\star + \xi_t$, where $\theta^\star \in \mathbb{R}^d$ is the underlying unknown parameter vector and $\{\xi_t\}_{t=0}^\infty$ is a sequence **of i.i.d.** zero-mean 1-sub-Gaussian random variables: $\mathbb{E}[e^{\lambda\xi_t}] \le e^{\frac{\lambda^2}{2}}$, $\forall\ \lambda \in \mathbb{R}$ (the constant 1 is without loss of generality). Hereafter, we shall call this model Model-C.

In the contextual bandit literature, an alternative model with a set of underlying unknown d-dimensional parameters $\{\theta_a^\star\}_{a\in[K]}$ is sometimes considered. In the alternative model, at time $t \in [T]$, the decision maker observes a d-dimensional context x_t, and if action $a \in [K]$ is chosen, the incurred regret is $r_{t,a} = x_t^\top \theta_a^\star + \xi_t$, where $\{\xi_t\}_{t=1}^\infty$ is similarly a sequence of i.i.d. zero-mean 1-sub-Gaussin random variables. We refer to this alternative model as Model-P.

Both models have been widely used in previous literature. For example, Model-C is adopted in Han et al. (2020) and Oh et al. (2021) and Model-P in Bastani et al. (2021)

and Bastani and Bayati (2020). The two models are in fact equivalent in the following sense: given Model-C, one can write $\tilde{x}_t = (x_{t,1}, \ldots, x_{t,K})$ and $\tilde{\theta}^*_a = (0, \ldots, \theta^*, \ldots, 0)$, and equivalently express $r_{t,a} = \tilde{x}_t^\top \tilde{\theta}^*_a + \xi_t$. Conversely, given Model-P, we can let $\tilde{x}_{t,a} = (0, \ldots, x_t, \ldots, 0)$ and $\tilde{\theta}^* = (\theta^*_1, \ldots, \theta^*_K)$. Then we have $r_{t,a} = \tilde{x}_{t,a}^\top \tilde{\theta}^* + \xi_t$. In this paper, we mainly focus on Model-C, while we shall also state parallel results under Model-P in Appendix F.

2.2. Assumptions

Without loss of generality (via normalization), we assume $\|\theta^*\|_2 \le 1$; the contexts $\{x_{t,a}\}_{a\in[K]}$ are random vectors i.i.d. drawn from a (Kd-dimensional) joint distribution each time: The independence is across time, but for each t, $x_{t,a}$s can be arbitrarily correlated across different as. We denote by a_t and r_{t,a_t} the (random) action chosen and the (random) reward incurred at time t: a_t is random because either it is randomly selected or the contexts $\{x_{t,a}\}_{a\in[K]}$ themselves are random, or both. We impose the following mild conditions on the context distribution.

Assumption 1 (Sub-Guassianity). *For $\forall a \in [K]$, the marginal distribution of $x_{t,a}$ is 1-sub-Gaussian, that is, $\mathbb{E}[X] = 0$ and $\mathbb{E}[\exp(v^\top X)] \le \exp(\|v\|^2/2)$, for $\forall v \in \mathbb{S}^{d-1}$.*

Remark 1. Because bounded contexts are automatically sub-Gaussian, this assumption is more general than the bounded contexts assumption commonly adopted in the contextual bandits literature (Wang et al. 2018, Kim and Paik 2019, Bastani and Bayati 2020, Bastani et al. 2021).

Assumption 2 (Diverse Covariate). *There are (possibly K-dependent) positive constants $\gamma(K)$ and $\rho(K)$, such that for any $\theta \in \mathbb{R}^d$ and any unit vector $v \in \mathbb{R}^d$, there is $\mathbb{P}((v^\top x_{t,a^*})^2 \ge \gamma(K)) \ge \rho(K)$, where $a^* = \arg\max_{a\in[K]} x_{t,a}^\top \theta$.*

Remark 2. The previous assumption ensures there is sufficient explaration even with a greedy algorithm (it is also the key condition used in Han et al. (2020) for the greedy algorithm there). We shall provide a thorough discussion on sufficient conditions for Assumption 2 in Section 2.3.

In low dimensions (Auer 2002, Chu et al. 2011), regret bounds of $\tilde{\Theta}(\sqrt{dT})$—which are minimax optimal up to log factors—have been obtained under upper confidence bound based algorithms such as LinREL in Auer (2002) or LinUCB in Chu et al. (2011). However, these algorithms and their Thompson sampling counterpart LinTS in Agrawal and Goyal (2013b) (which performs well empirically but often exhibit slightly worse regret bounds) cease to be effective in the high-dimensional regime as mentioned in the introduction. Of course, it's important to point out that absence of any further structure, $\tilde{\Theta}(\sqrt{dT})$ is the optimal regret bound and hence the best one can hope for even when d is very large. In this paper, we tackle this problem in the presence of sparsity, where only a few covariates influence rewards despite a large number of ambient covariates. In particular, we study the linear contextual bandits problem in the high-dimensional sparse regime: high-dimensional in the sense that d is large compared with T (the number of samples available in the entire learning horizon is small compared with the context dimension) and sparse in the sense that the underlying linear model is sparse: $\|\theta^*\|_0 \ll d$. We quantify them next.

Assumption 3 (Sparsity in High Dimension). *The dimension $d = \mathbf{Poly}(T)$ with sparse parameters: there exists some $\varepsilon > 0$ such that $\|\theta^*\|_0 \le s_0 = O(T^{1-\varepsilon})$.*

Remark 3. In statistical learning, a regime is considered high-dimensional if the dimension of the model is larger than the number of samples (Wainwright 2019). In our setting, this would translate to $d > T$. Consequently, our assumption that d can be any polynomial of T covers very high-dimensional regimes. Furthermore, learning becomes infeasible when d becomes even larger to, say, exponential in T, because a $\log d$ factor is present in the estimation accuracy even in the simple i.i.d. supervised learning setting (Hastie et al. 2015), which translates to a linear dependence on T. The sparsity requirement formalizes the precise requirement of $\|\theta^*\|_0 \ll d$. One can view $\|\theta^*\|_0$ (or its upper bound s_0) as the "intrinsic dimension" of the linear contextual bandits; consequently, s_0 should certainly be sublinear in T for learning to be effective. A typical regime of sparsity in statistical learning is $s_0 = O(\log d)$ (Wainwright 2019), which certainly meets the $s_0 = O(T^{1-\epsilon})$ requirement because $d = \mathbf{Poly}(T)$. Finally, in the previous assumption, we posit that an upper bound s_0 on the sparsity level is known to the decision maker. This assumption is standard and adopted for all the existing high-dimensional sparse linear contextual bandits (Wang et al. 2018, Kim and Paik 2019, Bastani and Bayati 2020) in their algorithm designs.

Finally, we work in the regime where the action set size K is not too large.

Assumption 4 (Not Many Actions). *The number of actions K satisfy the following two upper bounds: $\frac{\log K}{\gamma(K)\rho(K)} = O(d/s_0)$ and $\frac{\log K}{\gamma(K)\rho^3(K)} = O(\sqrt{T^{1-\beta}/s_0})$ for some $\beta > 0$.*

In our motivating applications, K is small (e.g., a constant number of actions) and easily satisfies this requirement, although this assumption can tolerate a much larger number of actions because $s_0 \ll d$. In practice, this regime typically suffices unless the number of actions is combinatorially large or when the action set is continuous, which would require a separate treatment.

2.3. Covariate Diversity Condition

In this section, we expand on Assumption 2 and provide a list of sufficient conditions for it.

Lemma 1. *The following are sufficient conditions for Assumption 2.*

1. *If for each* $a \in [K]$, $x_{t,a} \sim \mathcal{N}(0, \Sigma)$ *marginally, where* $\lambda_{\min}(\Sigma) > 0$, *then Assumption 2 holds with* $\gamma(K) = \frac{\lambda_{\min}(\Sigma)}{16}$ *and* $\rho(K) = \frac{1}{10}$.

2. *If there exists constants* $\alpha, c > 0$ *such that for each* $a \in [K]$ *and any unit vector* $v \in \mathbb{R}^d$,

$$\mathbb{E}[\exp(-\delta \cdot (v^\top x_{t,a})^2)] \le c \cdot \delta^{-\alpha}, \tag{1}$$

for any $\delta > 0$, then $\gamma(K) = \frac{\alpha}{e} \cdot (2cK)^{-1/\alpha}$ and $\rho(K) = \frac{1}{2}$.

3. *If there exists a constant* $\Lambda > 0$, *such that for each* $a \in [K]$ *and any unit vector* $v \in \mathbb{R}^d$, $v^\top \mathbb{E}[x_{t,a} x_{t,a}^\top] v \ge \Lambda$ *and* $\mathrm{Var}((v^\top x_{t,a})^2) \le \frac{\Lambda^2}{8K}$, *then* $\gamma(K) = \frac{\Lambda}{2}$ *and* $\rho(K) = \frac{1}{2}$.

4. *When* $K = 2$, *if there exists a constant* $\Lambda > 0$ *such that for any* $a \in [K]$, *any unit vector* $v \in \mathbb{R}^d$, $v^\top \mathbb{E}[x_{t,a} x_{t,a}^\top] v \ge \Lambda$, *and if there exists a constant* $\nu > 0$ *such that the joint distribution of* $(x_{t,1}, x_{t,2})$ *satisfies* $p(x_{t,1}, x_{t,2}) \ge \nu \cdot p(-x_{t,1}, -x_{t,2})$, *then* $\gamma(K) = \frac{\Lambda}{2}$ *and* $\rho(K) = \frac{\nu \Lambda^2}{64}$.

5. *When* $K > 2$, *suppose the following three conditions hold*:

   (a) *There exists a constant* $\Lambda > 0$, *such that for any* $a \in [K]$, *any unit vector* $v \in \mathbb{R}^d$, $v^\top \mathbb{E}[x_{t,a} x_{t,a}^\top] v \ge \Lambda$;

   (b) *There exists a constant* $\nu_1 > 0$ *such that the joint distribution of* $(x_{t,1}, \ldots, x_{t,K})$ *satisfies* $p(x_{t,1}, \ldots, x_{t,K}) \ge \nu_1 \cdot p(-x_{t,1}, \ldots, -x_{t,K})$;

   (c) *There exists a (possibly K-dependent) constant* $\nu_2(K) > 0$ *such that for any* $\theta \in \mathbb{R}^d$, *any permutation of* $[K]$ *denoted by* $\{\pi_1, \ldots, \pi_K\}$ *and any unit vector* $v \in \mathbb{R}^d$, *we have for any* $a \in [K]$

$$\begin{aligned}
&\nu_2(K) \cdot \mathbb{P}\left((v^\top x_{t,\pi_a})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta \right) \\
&\le \mathbb{P}\left((v^\top x_{t,\pi_1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta \right) \\
&\quad + \mathbb{P}\left((v^\top x_{t,\pi_K})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta \right).
\end{aligned}$$

Then Assumption 2 holds with $\gamma(K) = \frac{\Lambda}{2}$ *and* $\rho(K) = \frac{\nu_1 \nu_2(K) \Lambda^2}{128}$.

The proof of Lemma 1 can be found in Appendix B. Broadly speaking, the previous sufficient conditions can be categorized into two groups: Conditions 1–3 are assumptions on the marginal distribution of $x_{t,a}$, whereas Conditions 4 and 5 are on the joint distribution of $\{x_{t,a}\}_{a \in [K]}$. More specifically, Condition 1 is adopted from Han et al. (2020) and uses the property of the Gaussian distribution; Condition 2 characterizes a large class of distributions: In particular, if there exists a constant $\zeta > 0$, such that for any unit vector v and any $a \in [K]$, the distribution of $v^\top x_{t,a}$ is bounded by ζ, then this condition holds with $\alpha = 1/2$ and $c = \zeta\sqrt{\pi}/2$. The bounded density condition is similarly considered in Li et al. (2021) and is quite flexible: For example, when the coordinates $x_{t,a,j}$ are mutually independent across j, and the density of $x_{t,a,j}$ is bounded by ζ, the density of $v^\top x_{t,a}$ is bounded by $\sqrt{2}\zeta$ for any unit vector v (Rudelson and Vershynin 2015, theorem 1.2). Condition 3 requires the population covariance matrix of $x_{t,a}$ to be well conditioned, and the variance of $(v^\top x_{t,a})^2$ to be relatively small. Condition 4 is inspired by the diversity condition considered in Bastani et al. (2021, lemma 1) and Condition 5 by Oh et al. (2021, assumption 6).

These assumptions are on the covariates (as opposed to those on the underlying model), which are always testable (the covariates $\{x_{t,a}\}_{a \in [K]}$ can be fully observed). This fact is particularly appealing to practitioners.

2.4. Dynamic Batch Learning

In the standard online learning setting, the decision maker immediately observes the reward r_{t,a_t} after selecting action a_t at time t. After observing r_{t,a_t}, the decision maker can immediately incorporate this information in adapting her decision for action-selection at $t+1$. In particular, the decision maker can use all the historical information—including contexts $\{x_{\tau,a}\}_{\tau \le t, a \in [K]}$ and rewards $\{r_{\tau,a_\tau}\}_{\tau \le t-1}$—in deciding what action a_t to take at current time t.

In contrast, we consider a *dynamic batch learning* setting, where the decision maker is only allowed to partition the T units into (at most) M batches, and the reward corresponding to each unit in a batch can only be observed at the end of the batch. The decision maker can provision the partition *dynamically*: The decision maker can decide on how large the next batch is based on what has been observed in all previous batches, which includes all the contexts, the selected actions, and the corresponding rewards. The initial batch size is chosen without observing anything.

Formalizing the previous statement, given a maximum number of batches M, a dynamic batch learning algorithm $\mathbf{Alg} = (\mathcal{T}, \pi)$ has the following two components:

1. A *dynamic grid* $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$, with $0 = t_0 < t_1 < \cdots < t_M = T$, where each t_i is dynamically chosen at the end of t_{i-1} based on all the historical information available up to and including t_{i-1}. More specifically, prior to starting the decision making process, the decision maker decides on t_1, which indicates the length of the first batch. Having selected actions for each time in the first batch, the decision maker observes all the corresponding rewards at the end of t_1. Based on such information—including $\{a_t\}_{t=1}^{t_1}, \{x_{t,1}, \ldots, x_{t,K}\}_{t=1}^{t_1}$ and $\{r_{t,a_t}\}_{t=1}^{t_1}$—the decision maker then decides on what t_2 is. This dynamic grid partitioning process continues, and the decision maker always selects where the next batch ends at the end of current batch.

2. A sequence of policies $\pi = (\pi_1, \pi_2, \ldots, \pi_T)$ such that each π_t can only use reward information from all the prior batches and the contexts that can be observed up to t. That is, for a given t, if it lies in the ith batch ($t_{i-1} < t \le t_i$), then the policy to be used at t can use all the observed rewards from $\tau = 1$ to $\tau = t_{i-1}$, all the

selected actions from $\tau=1$ to $\tau=t-1$ and all the contexts information from $\tau=1$ to $\tau=t$.

Remark 4. Two special cases of a dynamic batch learning algorithm are worth mentioning. First, when the grid is fixed in advance—a static $\mathcal{T}$ is chosen completely at the beginning and not adapted during the learning process—we obtain a static batch learning algorithm, which is the class of algorithms considered in Han et al. (2020). Second, a further special case is the fixed grid $\mathcal{T}=\{1,2,\dots,T\}$ (i.e., $M=T$). This corresponds to the standard online learning setting where the decision maker need not select a grid. We also point out that $M=1$ is the other end of the spectrum, where no adaptation is allowed. In this case, irrespective of what one does, worst-case regret is always linear in T and regret (as defined next in Definition 1) is a meaningless (and thus the wrong) metric. Instead, one should adopt an offline learning viewpoint and adopt generalization error as the metric. This (offline learning in contextual bandits) would be an entirely new topic, and it has been well studied by a growing literature (see Zhao et al. (2014), Swaminathan and Joachims (2015), Joachims et al. (2018), Kitagawa and Tetenov (2018), and Kallus and Zhou (2018), and references therein).

To measure the performance of a dynamic batch learning algorithm **Alg**, we compare the cumulative reward obtained by **Alg** to that obtained by an **optimal** policy (an oracle that knows $\theta^\star$). This is formalized by regret, as defined next.

Definition 1. Let $\mathbf{Alg}=(\mathcal{T},\pi)$ be a dynamic batch learning algorithm. The regret of **Alg** is:

$$R_T(\mathbf{Alg}) \triangleq \sum_{t=1}^{T}\left(\max_{a\in[K]} x_{t,a}^\top\theta^\star - x_{t,a_t}^\top\theta^\star\right), \tag{2}$$

where $a_1,a_2,\dots,a_T$ are actions generated by **Alg** in the online decision-making process.

Remark 5. The regret defined previously is the same as used in standard online learning, but the feedback in our setting is much more restricted: Batches induce delays in obtaining reward feedback, and hence the decision maker cannot immediately incorporate the feedback into his subsequent decision making process. Consequently, all else equal, the regret will be a priori much larger when the decision maker is constrained to work with only a small number of batches.

3. Fundamental Limits: Regret Lower Bound

In this section, we present the minimax regret lower bound that characterizes the fundamental learning limits of dynamic batch learning in high-dimensional sparse linear contextual bandits.

Theorem 1. *Fix any s_0,d and T. Let $K=\log(T/s_0)$ and consider the problem $x_{t,a}\sim\mathcal{N}(0,I_d)$, $\forall a\in[K]$, $\forall t\in[T]$, where the contexts are independence across t. Then for any $M\le T$ and any dynamic batch learning algorithm **Alg**, we have*

$$\sup_{\theta^\star:\|\theta^\star\|_2\le 1,\|\theta^\star\|_0\le s_0}\mathbb{E}_{\theta^\star}[R_T(\mathbf{Alg})] \ge c\cdot\max\left(M^{-4}2^{-7M/2}\cdot\sqrt{Ts_0}\cdot\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},\sqrt{Ts_0}\right), \tag{3}$$

where $\mathbb{E}_{\theta^\star}$ denotes taking expectation w.r.t. the distribution based on the parameter $\theta^\star$, and $c>0$ is a numerical constant independent of (T,M,d,s_0).

We shall present the main steps in the proof Theorem 1 here and defer the details to Appendix C.

Remark 6. There are two terms on the right-hand side of Equation (3): The first term characterizes the dependence on M and the second term corresponds to the regret lower bound for the standard online learning setting. We have mentioned in the previous section that standard online learning (corresponding to $M=T$) is a simple special case of dynamic batch learning. Because a larger M provides better opportunities for adapting the decision-making process, a dynamic batch learning problem will only have worse regret compared with standard online learning setting. Consequently, a lower bound to standard online learning is immediately a lower bound to dynamic batch learning. Of course, the lower bound to dynamic batch learning will get worse, particularly when M is small (corresponding to limited chances to adapt one's decisions), and hence the first term on the right-hand side of Equation (3). We see that the break-even point—where the two lower bound terms equalize (up to log factors)—occurs at $M=O(\log\log(T/s_0))$. Consequently, taking into account the log terms, we see that when $M<O(\log\log(T/s_0))$, the first term dominates the lower bound, whereas the second term dominates the lower bound once M gets larger than $\Theta(\log\log(T/s_0))$.

Remark 7. In Theorem 1, the example used to show the lower bound satisfies Assumption 2 with $\gamma(K)=\rho(K)=O(1)$ and also satisfies Assumption 1. Furthermore, because the lower bound is established for any (s_0,d,T), it obviously holds for the regime given in Assumption 3 (because taking the supremum in a bigger set only results in a no-smaller lower bound). If in addition $s_0\log\log\left(\frac{T}{s_0}\right)=O(d)$ (a regime where d is slightly larger than s_0), then Assumption 4 is also satisfied for the problem construction in Theorem 1. Consequently, the lower bound holds under all four assumptions, under which the upper bound is subsequently established to match

the lower bound (up to log factors). Additionally, when $s_0 = d$ (the standard low-dimensional regime), our lower bound still holds, hence providing a fundamental limit that is not known even in that important special case. We do point out that in the low-dimensional regime where $s_0 = d$, unless $K = O(1)$, Assumption 4 does not hold, in which case the subsequent upper bound does not apply. Of course, this is not an issue at all because Han et al. (2020) already provided an upper bound for the low-dimensional setting under static batch design and matches the dynamic batch lower bound here, thereby completing the picture that even in the low-dimensional case and even when dynamic batch is allowed, one cannot do better than the static batch learning characterized in Han et al. (2020).

3.0.1. Main Proof Outline of Theorem 1. A key difficulty of the proof is that the grid is determined adaptively based on the observations from the previous batches. We briefly highlight the main proof steps here, each of which will be elaborated and rigorously formalized in a subsequent section.

We start from the regime of small M. Suppose $M = O(\log\log(T/s_0))$. Define for any $m \in [M]$,

$$T_m = \left\lfloor s_0 \cdot \left(\frac{T}{s_0}\right)^{\frac{1-2^{-m}}{1-2^{-M}}} \right\rfloor, \quad \Delta_m = \frac{1}{24 \cdot M^2 \cdot 2^{3M}} \cdot \left(\frac{T}{s_0}\right)^{-\frac{1-2^{1-m}}{2(1-2^{-M})}}.$$

Considering $K = 2^M$ arms, we shall construct a prior Q for θ^* and examine the regret under Q. Here, the prior is carefully designed such that for any $m \in [M]$, we can divide the 2^M arms into 2^{M-1} pairs such that the diffrence between each pair of arms is approximately the scale of Δ_m; the values of T_m and Δ_m are chosen such that the number of observations up to T_{m-1} is simply too few for the decision maker to distinguish the two arms in a pair (and learn a effective policy). Consequently, when the decision maker deploys this (ineffective) policy to this batch (from $T_{m-1}+1$ to T_m), even when restricted to the portion from $T_{m-1}+1$ to T_m, the total expected regret incurred—$(T_m - T_{m-1}) \cdot \Delta_m$—is still large. Section 3.1 details the construction of the prior, and Section 3.2 connects the worst-case regret to that under Q.

Given an **Alg**, of course its grid design $\{t_1, \ldots, t_M\}$ can be different from our "ideal" design $\{T_1, \ldots, T_M\}$. However, we now define for each $m \in [M]$ the "bad" event $B_m = \{t_{m-1} \le T_{m-1} < T_m \le t_m\}$: B_m is a "bad" event because, when B_m occurs, the number of observations up to t_{m-1} is too few (because $t_{m-1} \le T_{m-1}$) to distinguish pairs of arms that are Δ_m apart and learn an effective policy; when this (ineffective) policy is applied to this batch (from $t_{m-1}+1$ to t_m), the total expected regret incurred is still large (because $t_m \ge T_m$). In fact, we do not need a bad event to happen surely to guarantee that the total expected regret incurred is large: a bad event need only happen with a large enough probability to meet this purpose (with the probability taken over the randomness of the observations and the that of the parameters θ^*). Section 3.3 formalizes and establishes this step: If at least one B_m occurs with a large enough probability, then we obtain the desired final regret lower bound.

Section 3.4 is devoted to establishing that "if" is true. By a simple combinatorial argument, at least one of the B_m events will happen (under the convention that $t_0 = 0$, and since $t_M = T$, we are throwing $M-1$ points $t_1, t_2, \ldots, t_{M-1}$ into the M intervals partitioned by 0, $T_1, T_2, \ldots, T_{M-1}, T$, and hence the conclusion). In other words, $\{B_m\}_{m\in[M]}$ constitute a (nondisjoint) partition of the whole probability space. Hence, at least one bad event will happen with probability greater than $1/M$.

Finally, Section 3.5 establishes the lower bound for standard (fully) online learning ($M = T$): Because $M \le T$ in dynamic batch learning, this is clearly always a lower bound to the regret, which corresponds to the second term of the right-hand side of Equation (3). Taken together, these three steps complete the picture. We next dive into more details and begin with some useful notation.

3.1. Construction of the Prior

Let $\tilde{s}_0 = \lfloor s_0 \cdot 2^{-M} \rfloor \cdot 2^M$, and we have $s_0 - \tilde{s}_0 \le 2^M = O(\log(T/s_0))$. Next, we divide $[\tilde{s}_0]$ into consecutive subgroups at different levels of "resolution." At the first level of resolution, we divide $[\tilde{s}_0]$ into two consecutive group of equal sizes, denoted by I_0 and I_1, respectively, where

$$I_0 = \left\{1, \ldots, \frac{1}{2}\tilde{s}_0\right\}, \quad I_1 = \left\{\frac{1}{2}\tilde{s}_0 + 1, \ldots, \tilde{s}_0\right\};$$

at the second level of resolution, we further divide I_0 into two equal subgroups I_{00} and I_{01}, and I_1 into I_{10} and I_{11}, where

$$I_{00} = \left\{1, \ldots, \frac{1}{4}\tilde{s}_0\right\}, \quad I_{01} = \left\{\frac{1}{4}\tilde{s}_0 + 1, \ldots, \frac{1}{2}\tilde{s}_0\right\},$$
$$I_{10} = \left\{\frac{1}{2}\tilde{s}_0 + 1, \ldots, \frac{3}{4}\tilde{s}_0\right\}, \quad I_{11} = \left\{\frac{3}{4}\tilde{s}_0 + 1, \ldots, \tilde{s}_0\right\}.$$

Repeating the previous steps, at the Mth level of resolution we obtain 2^M subgroups of equal sizes:

$$I_{0\cdots00} = \left\{1, \ldots, \frac{1}{2^M}\tilde{s}_0\right\}, I_{0\cdots01} = \left\{\frac{1}{2^M}\tilde{s}_0 + 1, \ldots, \frac{1}{2^{M-1}}\tilde{s}_0\right\}, \cdots,$$
$$I_{1\cdots11} = \left\{\tilde{s}_0 - \frac{1}{2^M}\tilde{s}_0 + 1, \ldots, \tilde{s}_0\right\}.$$

To summarize, for any $m \in [M]$, a subgroup at the mth level of resolution is represented by a m-dimensional vector σ in $\Pi(m) := \{0,1\}^m$.

Next, we construct the prior Q on the true parameter θ. Generate $\theta_1, \ldots, \theta_M$ independently from $\mathsf{Unif}(\mathbb{S}^{\frac{\tilde{s}_0}{2^M}})$. For each $m \in [M]$, we define $\tilde{\theta}_m \in \mathbb{R}^{\tilde{s}_0}$ in the following

way: for each $\sigma \in \Pi(M)$,

$$\tilde{\theta}_m(I_\sigma) = \begin{cases} \theta_m & \sigma_m = 0, \\ -\theta_m & \sigma_m = 1, \end{cases}$$

where σ_m refers to the mth coordinate of σ. As a concrete example, for $m = 1$ and 2,

$$\tilde{\theta}_1 = (\underbrace{\theta_1, \ldots, \theta_1}_{2^{M-1} \text{ items}}, \underbrace{-\theta_1, \ldots, -\theta_1}_{2^{M-1} \text{ items}})$$

$$\tilde{\theta}_2 = (\underbrace{\theta_2, \ldots, \theta_2}_{2^{M-2} \text{ items}}, \underbrace{-\theta_2, \ldots, -\theta_2}_{2^{M-2} \text{ items}}, \underbrace{\theta_2, \ldots, \theta_2}_{2^{M-2} \text{ items}}, \underbrace{-\theta_2, \ldots, -\theta_2}_{2^{M-2} \text{ items}}).$$

Setting $\tilde{\theta} = \sum_{m=1}^{M} \Delta_m \theta_m$, we construct $\theta := f(\theta_1, \theta_2, \ldots, \theta_M) \in \mathbb{R}^d$ by letting its first $\tilde{s}_0$ coordinates be $\tilde{\theta}$ and the others zero. It can checked that $\|\theta\|_2 = \|\tilde{\theta}\|_2 \le \sum_{m=1}^{M} \Delta_m \|\tilde{\theta}_m\|_2 \le 1$ and $\|\theta\|_1 = \tilde{s}_0$.

We now proceed to specify the joint distribution of the $K = 2^M$ arms. For each $t \in [T]$, we first draw $x_t \sim \mathcal{N}(0, I_d)$. To simplify the notation, we let $S = \{1, 2, \ldots, \tilde{s}_0\}$ and $S^c = \{\tilde{s}_0 + 1, \ldots, d\}$. For each $a \in [K]$, we first let $x_{t,a}(S^c) = x_t(S^c)$. It remains to specify the first $\tilde{s}_0$ coordinates of $x_{t,a}$. To do so, we again divide S into 2^M consecutive groups, each represented by $\sigma \in \Pi(M)$, and will specify the value of each group. Given an arm a, we can uniquely write $a = 1 + \sum_{m=1}^{M} a_m \cdot 2^{m-1}$ where $a_m \in \{0, 1\}$ for each $m \in [M]$; define a mapping $\mathcal{M}_a : \Pi(M) \mapsto \Pi(M)$, where for any $m \in [M]$, $\mathcal{M}_a(\sigma)_m = (1 - a_m) \cdot \sigma_m + a_m \cdot (1 - \sigma_m)$. We then let $x_{t,a}(\sigma) = x_t(\mathcal{M}_a(\sigma))$, for any $\sigma \in \Pi(M)$. For example, when $M = 2$, we have four arms, where

$$x_{t,1} = x_t,\ x_{t,2} = (x_t(I_1), x_t(I_0), x_t(S^c)),$$
$$x_{t,3} = (x_t(I_{01}), x_t(I_{00}), x_t(I_{11}), x_t(I_{10}), x_t(S^c)),$$
$$x_{t,4} = (x_t(I_{11}), x_t(I_{10}), x_t(I_{01}), x_t(I_{00}), x_t(S^c)).$$

By construction, for any $a \in [K]$, $x_{t,a} \sim \mathcal{N}(0, I_d)$ marginally, thus satisfying Assumption 2.

3.2. Notation for Regret Decomposition

We streamline the notation for a regret decomposition that will be used throughout:

$$\sup_{\theta^*: \|\theta^*\|_2 \le 1, \|\theta^*\|_0 \le s_0} \mathbb{E}_{\theta^*}[R_T(\mathbf{Alg})] \ge \mathbb{E}_Q \mathbb{E}_\theta[R_T(\mathbf{Alg})]$$
$$= \sum_{t=1}^{T} \mathbb{E}_Q \left(\mathbb{E}_x \mathbb{E}_{P^t_{\theta,x}} \left[\max_{a \in [K]} x_{t,a}^\top \theta - x_{t,a_t}^\top \theta \right] \right),$$

where $\mathbb{E}_Q$ denotes taking expectation with respect to the prior Q of θ, $\mathbb{E}_x$ denotes taking expectation with respect to all the random contexts at all times (note that it is both equivalent and conceptually simpler to imagine all the contexts $x = \{x_{t,a}\}_{t \in [T], a \in [K]}$ have been drawn once for all ahead of time before the decision-making process starts), and $P^t_{\theta,x}$ denotes the distribution of all observed rewards before time t (and hence before the start of the current batch that contains t) conditional on the parameter θ and the contexts x. Per its definition, the distributions $P^t_{\theta,x}$ and $P^{t+1}_{\theta,x}$ are the same if t and $t+1$ belong to the same batch.

Recall that for each $j \in [2^M]$, we write $j = 1 + \sum_{m=1}^{M} j_m \cdot 2^{m-1}$. Then for each $t \in [T]$ and any $m \in [M]$,

$$\max_{a \in [K]} (x_{t,a}^\top \theta - x_{t,a_t}^\top \theta) = \sum_{j \in [K]} \mathbf{1}\{a_t = j\} \cdot \max_{a \in [K]} (x_{t,a}^\top \theta - x_{t,j}^\top \theta)$$
$$\overset{(a)}{=} \sum_{j \in [K]: j_m = 0} \mathbf{1}\{a_t = j\} \cdot \max_{a \in [K]} (x_{t,a}^\top \theta - x_{t,j}^\top \theta) + \mathbf{1}\{a_t = j + 2^{m-1}\} \cdot \max_{a \in [K]} (x_{t,a}^\top \theta - x_{t,j+2^{m-1}}^\top \theta)$$
$$\ge \sum_{j \in [K]: j_m = 0} \mathbf{1}\{a_t = j\} \cdot \max_{a \in \{j, j+2^{m-1}\}} (x_{t,a}^\top \theta - x_{t,j}^\top \theta) + \mathbf{1}\{a_t = j + 2^{m-1}\} \cdot \max_{a \in \{j, j+2^{m-1}\}} (x_{t,a}^\top \theta - x_{t,j+2^{m-1}}^\top \theta)$$
$$= \sum_{j \in [K]: j_m = 0} \mathbf{1}\{a_t = j\} \cdot (x_{t,j+2^{m-1}}^\top \theta - x_{t,j}^\top \theta)_+ + \mathbf{1}\{a_t = j + 2^{m-1}\} \cdot (x_{t,j+2^{m-1}}^\top \theta - x_{t,j}^\top \theta)_-, \tag{4}$$

where in step (a) we categorize the arms into two groups by the value of j_m. For a j such that $j_m = 0$, we can write

$$x_{t,j+2^{m-1}}^\top \theta - x_{t,j}^\top \theta = \sum_{\sigma \in \Pi(M)} x_{t,j+2^{m-1}}(I_\sigma)^\top \theta(I_\sigma) - x_{t,j}(I_\sigma)^\top \theta(I_\sigma)$$
$$= \sum_{\sigma \in \Pi(M): \sigma_m = 0} x_{t,j+2^{m-1}}(I_\sigma)^\top \theta(I_\sigma) - x_{t,j}(I_\sigma)^\top \theta(I_\sigma) + \sum_{\sigma \in \Pi(M): \sigma_m = 1} x_{t,j+2^{m-1}}(I_\sigma)^\top \theta(I_\sigma) - x_{t,j}(I_\sigma)^\top \theta(I_\sigma)$$
$$= 2\Delta_m \cdot \theta_m^\top \left(\sum_{\sigma \in \Pi(M): \sigma_m = 1} x_t(I_\sigma) - \sum_{\sigma \in \Pi(M): \sigma_m = 0} x_t(I_\sigma) \right).$$

To simplify the notation, we define

$$d_{m,t} = \sum_{\sigma \in \Pi(M): \sigma_m = 2} x_t(I_\sigma) - \sum_{\sigma \in \Pi(M): \sigma_m = 1} x_t(I_\sigma), \quad u_{m,t} = \frac{d_{m,t}}{\|d_{m,t}\|_2},$$

and $\mathcal{A}_m = \{j \in [K] : j_m = 0\}$. With the previous expressions, we continue decomposing the regret

$$(4) = 2\Delta_m \sum_{j \in \mathcal{A}_m} \mathbf{1}\{a_t = j\} \cdot (d_{m,t}^\top \theta_m)_+ + \mathbf{1}\{a_t = j + 2^{m-1}\} \cdot (d_{m,t}^\top \theta_m)_-$$
$$= 2\Delta_m \cdot \mathbf{1}\{j \in \mathcal{A}_m\} \cdot (d_{m,t}^\top \theta_m)_+ + \mathbf{1}\{j \in \mathcal{A}_m^c\} \cdot (d_{m,t}^\top \theta_m)_-.$$

As a result, we have

$$\mathbb{E}_Q \mathbb{E}_{P^t_{\theta,x}} \left[\max_{a \in [K]} (x_{t,a}^\top \theta - x_{t,a_t}^\top \theta) \right]$$
$$\ge 2\Delta_m \cdot \mathbb{E}_Q [(d_{m,t}^\top \theta_m)_+ \cdot \mathbb{E}_{P^t_{\theta,x}}[\mathbf{1}\{a_t \in \mathcal{A}_m\}] + (d_{m,t}^\top \theta_m)_- \cdot \mathbb{E}_{P^t_{\theta,x}}[\mathbf{1}\{a_t \in \mathcal{A}_m^c\}]], \tag{5}$$

where we note that conditioned on θ and x, a_t depends on the distribution of observed rewards $P^t_{\theta,x}$ (hence the inner expectation is taken with respect to this distribution). Through a change of measure, we define two new

probability measures via

$$\frac{dQ^+_{m,t}}{dQ}(\theta)=\frac{(d^\top_{m,t}\theta_m)_+}{Z_m(d_{m,t})},\ \frac{dQ^-_{m,t}}{dQ}(\theta)=\frac{(d^\top_{m,t}\theta_m)_-}{Z_m(d_{m,t})},$$

where $Z_m(d_{m,t})=\mathbb{E}_Q[(d^\top_{m,t}\theta_m)_+]=\mathbb{E}_Q[(d^\top_{m,t}\theta_m)_-]$ is a common normalizing factor. Then,

$$\begin{aligned}&\mathbb{E}_Q\mathbb{E}_{P^t_{\theta,x}}\left[\max_{a\in[K]}(x^\top_{t,a}\theta-x^\top_{t,a_t}\theta)\right]\\&\geq 2\Delta_m Z_m(d_{m,t})\cdot\left(\mathbb{E}_{P^t_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{a_t\in\mathcal{A}_m\}]\right.\\&\quad\left.+\mathbb{E}_{P^t_{\theta,x}\circ Q^-_{m,t}}[\mathbf{1}\{a_t=\mathcal{A}^c_m\}]\right),\end{aligned}$$

where $P^t_{\theta,x}\circ Q^+_{m,t}$ (respectively, $P^t_{\theta,x}\circ Q^-_{m,t}$) is a mixed distribution: θ is drawn from $Q^+_{m,t}$ (respectively, $Q^-_{m,t}$) and observed rewards are then drawn from $P^t_{\theta,x}$. Note that $Z_m(\cdot)$ is a function and $Z_m(d_{m,t})$ emphasizes that the common normalizing factor depends on $d_{m,t}$.

Reparametrizing of the regret in terms of the two newly defined priors allows us to connect the regret with the distance between measures, from which lower bounds can be established with information-theoretic tools.

3.3. Regret Lower Bound When a "Bad" Event Happens with Large Probability

When a "bad" event B_m is likely to happen under prior Q, large regret follows.

Lemma 2. *If there exists $m\in[M]$, such that*

$$\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x\left[Z_m(d_{m,t})\cdot\mathbb{E}_{P_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]\right]\geq\frac{T_m-T_{m-1}}{8\cdot 2^{\frac{M}{2}}M^2},\tag{6}$$

then there eixsts a numerical constant $c>0$, independent of (T,M,d,s_0), such that,

$$\sup_{\theta^\star:\|\theta^\star\|_2\leq 1,\|\theta^\star\|_0\leq s_0}\mathbb{E}_{\theta^\star}[R_T(\mathbf{Alg})]\geq\frac{c}{M^4 2^{3M}}\sqrt{Ts_0}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}.$$

Using the decomposition of the reget, we have for any $m\in[M]$,

$$\begin{aligned}&\sup_{\theta^\star:\|\theta^\star\|_2\leq 1,\ \|\theta^\star\|_0\leq s_0}\mathbb{E}_{\theta^\star}[R_T(\mathbf{Alg})]\\&\geq 2\Delta_m\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x\left[Z_m(d_{m,t})\cdot\left(\mathbb{E}_{P^t_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{a_t\in\mathcal{A}_m\}]\right.\right.\\&\quad\left.\left.+\mathbb{E}_{P^t_{\theta,x}\circ Q^-_{m,t}}[\mathbf{1}\{a_t\in\mathcal{A}^c_m\}]\right)\right]\\&\overset{(a)}{\geq} 2\Delta_m\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x[Z_m(d_{m,t})\cdot(1-\mathsf{TV}(P^t_{\theta,x}\circ Q^+_{m,t},P^t_{\theta,x}\circ Q^-_{m,t}))]\\&\overset{(b)}{\geq} 2\Delta_m\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x[Z_m(d_{m,t})\cdot(1-\mathsf{TV}(P^{T_m}_{\theta,x}\circ Q^+_{m,t},P^{T_m}_{\theta,x}\circ Q^-_{m,t}))],\end{aligned}$$

where step (a) is because $P(A)+Q(A^c)\geq 1-\mathsf{TV}(P,Q)$, and step (b) follows from the data processing inequality of the total variation distance (Lemma A.2). For the total variation,

$$\begin{aligned}&1-\mathsf{TV}\left(P^{T_m}_{\theta,x}\circ Q^+_{m,t},P^{T_m}_{\theta,x}\circ Q^-_{m,t}\right)\\&=\int\min\left(dP^{T_m}_{\theta,x}\circ Q^+_{m,t},dP^{T_m}_{\theta,x}\circ Q^-_{m,t}\right)\\&\geq\int_{B_m}\min\left(dP^{T_m}_{\theta,x}\circ Q^+_{m,t},dP^{T_m}_{\theta,x}\circ Q^-_{m,t}\right)\\&=\frac{1}{2}\int_{B_m}\left(dP^{T_m}_{\theta,x}\circ Q^+_{m,t}+dP^{T_m}_{\theta,x}\circ Q^-_{m,t}\right.\\&\quad\left.-|dP^{T_m}_{\theta,x}\circ Q^+_{m,t}-dP^{T_m}_{\theta,x}\circ Q^-_{m,t}|\right)\\&=\frac{1}{2}\int_{B_m}\left(dP^{T_{m-1}}_{\theta,x}\circ Q^+_{m,t}+dP^{T_{m-1}}_{\theta,x}\circ Q^-_{m,t}\right.\\&\quad\left.-|dP^{T_{m-1}}_{\theta,x}\circ Q^+_{m,t}-dP^{T_{m-1}}_{\theta,x}\circ Q^-_{m,t}|\right),\end{aligned}\tag{7}$$

where the last equality uses the fact that on the event B_m, $dP^{T_{m-1}}_{\theta,x}=dP^{T_m}_{\theta,x}$. Using the property that $\mathsf{TV}(P,Q)=\frac{1}{2}\int|dP-dQ|$ and $|P(A)-Q(A)|\leq\mathsf{TV}(P,Q)$, we have

$$\begin{aligned}(7)&=\frac{1}{2}\left(\mathbb{E}_{P^{T_{m-1}}_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]+\mathbb{E}_{P^{T_{m-1}}_{\theta,x}\circ Q^-_{m,t}}[\mathbf{1}\{A_m\}]\right)\\&\quad-\mathsf{TV}\left(dP^{T_{m-1}}_{\theta,x}\circ Q^+_{m,t},dP^{T_{m-1}}_{\theta,x}\circ Q^-_{m,t}\right)\\&\geq\mathbb{E}_{P_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]-\frac{3}{2}\mathsf{TV}\left(P^{T_{m-1}}_{\theta,x}\circ Q^+_{m,t},P^{T_{m-1}}_{\theta,x}\circ Q^-_{m,t}\right).\end{aligned}$$

Applying Pinsker's inequality (Lemma A.3), we have

$$\begin{aligned}&\mathsf{TV}\left(P^{T_{m-1}}_{\theta,x}\circ Q^+_{m,t},P^{T_{m-1}}_{\theta,x}\circ Q^-_{m,t}\right)\\&\leq\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(P^{T_{m-1}}_{\theta,x}\circ Q^+_{m,t}\|P^{T_{m-1}}_{\theta,x}\circ Q^-_{m,t}\right)}.\end{aligned}\tag{8}$$

To simplify the right-hand side of Equation (8), we use the rotational invariance of the uniform distribution. First, let $v_1,v_2,\ldots,v_{2^{-M}\tilde{s}_0}$ be an orthonormal basis of $\mathbb{R}^{2^{-M}\tilde{s}_0}$, where $v_1=u_{m,t}$; define two rotational matrices $R_1=[v_1,v_2,\ldots,v_{2^{-M}\tilde{s}_0}]$ and $R_2=[-v_1,v_2,\ldots,v_{2^{-M}\tilde{s}_0}]$; letting $\theta'_m=\theta_m-2(v_1^\top\theta_m)v_1=R_1R_2^\top\theta_m$, we have $\theta'_m\overset{\mathrm{d}}{=}\theta_m$ and $\theta'^\top_m d_{m,t}=-\theta^\top_m d_{m,t}$. Furthermore, let θ' denote the parameter induced by $\theta_1,\ldots,\theta_{m'},\ldots,\theta_M$—that is, $\theta':=f(\theta_1,\ldots,\theta'_m,\ldots,\theta_M)$—we then have $\theta'\overset{\mathrm{d}}{=}\theta$ and

$$\begin{aligned}(8)&=\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(P^{T_{m-1}}_{\theta,x}\circ Q^+_{m,t}\|P^{T_{m-1}}_{\theta',x}\circ Q^+_{m,t}\right)}\\&\leq\sqrt{\frac{1}{2}\mathbb{E}_{Q^+_{m,t}}\left[D_{\mathrm{KL}}\left(P^{T_{m-1}}_{\theta,x}\|P^{T_{m-1}}_{\theta',x}\right)\right]},\end{aligned}$$

where the inequality is due to the joint convexity of the Kullback–Leibler (KL) divergence (Lemma A.4). The

KL divergence can then be explicitly computed:

$$\begin{aligned}
&\mathbb{E}_{Q^+_{m,t}}\left[D_{\mathrm{KL}}(P^{T_{m-1}}_{\theta,x}\|P^{T_{m-1}}_{\theta',x})\right]\\
&=\frac{1}{2}\sum_{\tau=1}^{T_{m-1}}\mathbb{E}_{Q^+_{m,t}}[(f(\theta_1,\dots,\theta_m,\dots,\theta_M)-f(\theta_1,\dots,\theta_{m'},\dots,\theta_M))^\top x_{\tau,a_\tau}]^2\\
&=2\Delta_m^2\cdot\mathbb{E}_{Q^+_{m,t}}[|u_{m,t}^\top\theta_m|^2]\cdot u_{m,t}^\top\left(\sum_{\tau=1}^{T_{m-1}}h_{\tau,a_\tau}h_{\tau,a_\tau}^\top\right)u_{m,t}\\
&\le 2\Delta_m^2\cdot\mathbb{E}_{Q^+_{m,t}}[|u_{m,t}^\top\theta_m|^2]\cdot u_{m,t}^\top\left(\sum_{\tau=1}^{T_{m-1}}\sum_{j\in[K]}h_{\tau,j}h_{\tau,j}^\top\right)u_{m,t},
\end{aligned}$$

where $h_{\tau,j}=\sum_{\sigma\in\Pi(M):\sigma_m=0}x_{\tau,j}(I_\sigma)-\sum_{\sigma\in\Pi(M):\sigma_m=1}x_{\tau,j}(I_\sigma)$. Note that

$$\begin{aligned}
\mathbb{E}_{Q^+_{m,t}}[|u_{m,t}^\top\theta_m|^2]&=\frac{\|d_{m,t}\|_2}{2Z_m(d_{m,t})}\cdot\mathbb{E}_Q[|u_{m,t}^\top\theta_m|^3]\\
&=\frac{\mathbb{E}_Q[|\theta_{m,1}|^3]}{2Z_m(u_{m,t})}=\frac{\mathbb{E}_Q[|\theta_{m,1}|^3]}{\mathbb{E}_Q[|\theta_{m,1}|]}\\
&\overset{(a)}{=}\frac{2}{2^{-M}\cdot\tilde{s}_0+1}\le\frac{2^{M+1}}{\tilde{s}_0},
\end{aligned}$$

where step (a) follows from Lemma E.1. We then can lower bound the regret as

$$\begin{aligned}
&\sup_{\theta^\star:\|\theta^\star\|_2\le1,\|\theta^\star\|_0\le s_0}\mathbb{E}_{\theta^\star}[R_T(\mathbf{Alg})]\\
&\ge 2\Delta_m\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x\left[Z_m(d_{m,t})\cdot\left(\mathbb{E}_{P_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]\right.\right.\\
&\qquad\left.\left.-\frac{3}{2}\sqrt{\frac{2^{M+1}\Delta_m^2}{\tilde{s}_0}\cdot u_{m,t}^\top\left(\sum_{\tau=1}^{T_{m-1}}\sum_{j\in[K]}h_{\tau,j}h_{\tau,j}^\top\right)u_{m,t}}\right)\right]\\
&\overset{(a)}{\ge} 2\Delta_m\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x\left[Z_m(d_{m,t})\cdot\left(\mathbb{E}_{P_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]\right.\right.\\
&\qquad\left.\left.-\frac{3}{2}\sqrt{\frac{2^{3M+1}\Delta_m^2T_{m-1}}{\tilde{s}_0}}\right)\right]\\
&\overset{(b)}{\ge} 2\Delta_m\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x\left[Z_m(d_{m,t})\cdot\left(\mathbb{E}_{P_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]-\frac{1}{2\cdot2^{M+2}M^2}\right)\right],
\end{aligned}\tag{9}$$

where step (a) uses the independence between $(x_{t,1},x_{t,2})$ and $\{(x_{\tau,1},x_{\tau,2})\}_{\tau\le T_{m-1}}$ and the concavity of $x\mapsto\sqrt{x}$; step (b) is due to the choice of Δ_m and T_{m-1}. Note also that

$$\begin{aligned}
\mathbb{E}_x[Z_m(d_{m,t})]&=\mathbb{E}_x\left[\frac{\|d_{m,t}\|_2}{2}\right]\cdot\mathbb{E}_Q[|u_t^\top\theta|]\\
&=\mathbb{E}_x\left[\frac{\|d_{m,t}\|_2}{2}\right]\cdot\mathbb{E}_Q[|\theta_1|]\overset{(a)}{\le}\frac{2^{\frac{M}{2}}}{\sqrt{\tilde{s}_0}}\mathbb{E}_x[\|d_{m,t}\|_2]\\
&\le 2^{\frac{M}{2}},
\end{aligned}$$

where step (a) follows from Lemma E.1. Consequently,

$$\begin{aligned}
(9)\ge 2\Delta_m\left(\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x[Z_m(d_{m,t})\cdot\right.\\
\left.\mathbb{E}_{P_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]]-\frac{1}{2\cdot2^{\frac{M}{2}+2}M^2}.\right)
\end{aligned}$$

Finally, letting m be the batch that satisfies Equation (6), we have

$$\begin{aligned}
\sup_{\theta^\star:\|\theta^\star\|_2\le1,\|\theta^\star\|_0\le s_0}\mathbb{E}_{\theta^\star}\cdot[R_T(\mathbf{Alg})]&\ge\frac{(T_m-T_{m-1})\Delta_m}{2\cdot2^{\frac{M}{2}+2}M^2}\\
&\ge\frac{c}{M^4 2^{7M/2}}\cdot\sqrt{s_0T}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}.
\end{aligned}$$

3.4. Bad Event Happens with Large Enough Probability

Our main result here is that a bad event occurs with sufficiently high probability that (6) holds.

Lemma 3. *There exists some $m\in[M]$, such that*

$$\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x[Z_m(d_{m,t})\cdot\mathbb{E}_{P_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]]\ge\frac{T_m-T_{m-1}}{2^{\frac{M}{2}+2}M^2}.$$

Because the union of $\{B_m\}_{m\in[M]}$ is the whole space, by a union bound, we have $\sum_{m=1}^M P(B_m)\ge P(\cup_{m=1}^M B_m)=1$, where P is any probability measure. Hence $P(B_m)\ge1/M$ for at least one m. For any $m\in[M]$,

$$\begin{aligned}
&\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x[Z_m(d_{m,t})\cdot\mathbb{E}_{P_{\theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{B_m\}]]\\
&=\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x\mathbb{E}_Q[(d_{m,t}^\top\theta_m)_+\cdot P_{\theta,x}(B_m)].
\end{aligned}$$

Conditional on $\{x_{t,a}\}_{t\le T_{m-1},a\in[K]}$, $\mathbf{1}\{B_m\}$ is independent of $\{x_{t,a}\}_{t>T_{m-1},a\in[K]}$. Hence,

$$\begin{aligned}
P_{\theta,x}(B_m)&=\mathbb{P}_\theta(t_{m-1}\le T_{m-1}<T_m\le t_m|\{x_{1,a}\}_{a\in[K]},\\
&\qquad\dots,\{x_{T,a}\}_{a\in[K]})\\
&=\mathbb{P}_\theta(t_{m-1}\le T_{m-1}<T_m\le t_m|\{x_{1,a}\}_{a\in[K]},\\
&\qquad\dots,\{x_{T_{m-1},a}\}_{a\in[K]}).
\end{aligned}$$

Consequently, using the independence between context x across t, we have

$$\begin{aligned}
&\mathbb{E}_x[\mathbb{E}_Q[(d_{m,t}^\top\theta_m)_+P_{\theta,x}(B_m)]]=\mathbb{E}_Q[\mathbb{E}_x[(d_{m,t}^\top\theta_m)_+P_{\theta,x}(B_m)]]\\
&=\mathbb{E}_Q[\mathbb{E}_x[(d_{m,t}^\top\theta_m)_+]\cdot\mathbb{E}_x[P_{\theta,x}(B_m)]]=\mathbb{E}_Q[\mathbb{E}_x[(d_{M,T}^\top\theta_m)_+]\\
&\quad\cdot\mathbb{E}_x[P_{\theta,x}(B_m)]]\\
&=\mathbb{E}_Q[\mathbb{E}_x[(d_{M,T}^\top\theta_m)_+P_{\theta,x}(B_m)]]=\mathbb{E}_x[\mathbb{E}_Q[(d_{M,T}^\top\theta_m)_+P_{\theta,x}(B_m)]].
\end{aligned}$$

Using the previous result, we obtain that

$$\begin{aligned}
&\sum_{t=T_{m-1}+1}^{T_m} \mathbb{E}_x\left[Z_m(d_{m,t}) \cdot \mathbb{E}_{P_{\theta,x}\circ Q_{1,m}^t}[\mathbf{1}\{B_m\}]\right] \\
&= \sum_{t=T_{m-1}+1}^{T_m} \mathbb{E}_x \mathbb{E}_Q\left[(d_{M,T}^\top \theta_m)_+ \cdot P_{\theta,x}(B_m)\right] \\
&\geq \sum_{t=T_{m-1}+1}^{T_m} \mathbb{E}_x \mathbb{E}_Q\left[\min_{m'\in[M]}\{(d_{M,T}^\top \theta_{m'})_+\} \cdot P_{\theta,x}(B_m)\right] \\
&\overset{(a)}{=} \sum_{t=T_{m-1}+1}^{T_m} \tilde{Z} \cdot \mathbb{E}_{P_{\theta,x}\circ\tilde{Q}}[\mathbf{1}\{B_m\}] \\
&= (T_m - T_{m-1})\tilde{Z} \cdot \mathbb{E}_{P_{\theta,x}\circ\tilde{Q}}[\mathbf{1}\{B_m\}],
\end{aligned}$$

where in step (a) we define the measure $\tilde{Q}$ via $\frac{d\tilde{Q}}{dQ\times dP_x}(x,\theta) = \frac{\min_{m'\in[M]}(d_{M,T}^\top \theta_{m'})_+}{\tilde{Z}}$, and $\tilde{Z} = \mathbb{E}_x\mathbb{E}_Q[\min_{m'\in[M]}(d_{M,T}^\top \theta_{m'})_+]$ is a normalizing constant.

$$\begin{aligned}
\tilde{Z} &= \mathbb{E}_x[\|d_{M,T}\|_2 \cdot \mathbb{E}_Q\left[\min_{m\in[M]} (u_{M,T}^\top \theta_m)_+\right]] \\
&= \mathbb{E}_x[\|d_{M,T}\|_2 \cdot \mathbb{E}_Q\left[\min_{m\in[M]} (\theta_{m,1})_+\right]],
\end{aligned}$$

where the last equality is due the fact that $\theta_1,\dots,\theta_M$ are independent of each other and the rotational invariance of the uniform distribution. Note also

$$\begin{aligned}
&\mathbb{E}_Q\left[\min_m (\theta_{m,1})_+\right] \\
&= \int_0^\infty \mathbb{P}\left(\min_m (\theta_{m,1})_+ > t\right)\mathrm{d}t = \int_0^\infty \mathbb{P}\left((\theta_{1,1})_+ > t\right)^M \mathrm{d}t \\
&= \frac{1}{2^M}\int_0^\infty \mathbb{P}(|\theta_{1,1}| > t)^M \mathrm{d}t = \frac{1}{2^M}\int_0^\infty \mathbb{P}(|\theta_{1,1}|^2 > t^2)^M \mathrm{d}t \\
&\geq \frac{1}{2^M}\int_0^{\mathsf{B}(\frac{1}{2},\frac{\tilde{s}_0 2^{-M}-1}{2})/2} \left(1 - \frac{2t}{\mathsf{B}(\frac{1}{2},\frac{\tilde{s}_0 2^{-M}-1}{2})}\right)^M \mathrm{d}t \\
&= \frac{\mathsf{B}(\frac{1}{2},\frac{\tilde{s}_0 2^{-M}-1}{2})}{2^{M+1}(M+1)} \geq \frac{1}{2^{\frac{M+1}{2}}(M+1)\sqrt{\tilde{s}_0}},
\end{aligned}$$

where $\mathsf{B}(\alpha,\beta)$ denotes the beta function with parameters α and β. With the previous result,

$$\tilde{Z} \geq \frac{1}{2^{\frac{M+1}{2}}\cdot(M+1)}.$$

Because $\sum_{m=1}^M \mathbb{E}_{P_{\theta,x}\circ\tilde{Q}}[\mathbf{1}\{B_m\}] \geq 1$, there exists $m\in[M]$, such that $\mathbb{E}_{P_{\theta,x}\circ\tilde{Q}}[\mathbf{1}\{B_m\}] \geq \frac{1}{M}$, and hence,

$$\tilde{Z}\cdot\mathbb{E}_{P_{\theta,x}\circ\tilde{Q}}[\mathbf{1}\{B_m\}] \geq \frac{\tilde{Z}}{M} \geq \frac{1}{2^{\frac{M}{2}+2}M^2}.$$

Finally for this m, $\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x[Z_m(d_{m,t})\cdot\mathbb{E}_{P_{\theta,x}\circ Q_{m,t}^+}[\mathbf{1}\{B_m\}]] \geq \frac{(T_m-T_{m-1})}{2^{\frac{M}{2}+2}M^2}$.

3.5. Lower Bound for Fully Online Learning Setting

Thus far, we have established the left-hand side of (3) is greater or equal to the first term on the right-hand side when $M = O(\log\log(T/s_0))$. When $M = \Omega(\log\log(T/s_0))$, the first term is dominated by the second term, so it suffices to show that the regret is lower bounded by the second term. Lemma 4 completes the picture by showing the second part of the inequality.

Lemma 4. *When $M = T$, there exists a two-arm setting with independent Guassian contexts, for which we have (for some numerical constant c independent of T, M, d, s_0):*

$$\sup_{\theta^\star:\|\theta^\star\|_2\leq 1, \|\theta^\star\|_0\leq s_0} \mathbb{E}_{\theta^\star}[R_T(\mathbf{Alg})] \geq c\cdot\sqrt{Ts_0}.$$

The proof is a simple variant of the first part, and the proof is in Appendix C.1. Our online regret lower bound recovers the lower bound obtained in Chu et al. (2011)—Their lower bound is stated in the dense and low-dimensional setting, but the adaptation is straightforward.

4. Achievable Guarantees: Regret Upper Bound

In this section, we propose the *LASSO batch greedy learning* (LBGL) algorithm, similar in spirit to the (low-dimensional) greedy bandit algorithm (Bastani et al. 2021), to tackle the high-dimensional dynamic batch learning problem. This simple algorithm is minimax optimal (up to log factors).

4.1. LASSO Batch Greedy Learning

LBGL has two important features: (1) at each time t, it exploits the current estimate of the true parameter $\theta^\star$ without further exploration; and (2) it uses a static grid that is not adaptive (of course, a static grid is a particular type of dynamic grid). As it turns out, this already achieves the optimal regret bound. Concretely, given a grid choice $\mathcal{T} = \{t_1,\dots,t_M\}$, at the beginning of batch m, the algorithm constructs a Lasso estimate $\hat{\theta}_{m-1}$ of the true parameter using the data in the previous batches; then it selects the action $a\in[K]$ that maximizes the estimated reward $x_{t,a}^\top\hat{\theta}_{m-1}$ for any $t\in\{t_{m-1}+1,\dots,t_m\}$; at the end of the mth batch, the algorithm updates the estimate of the underlying parameters with the new observations in the current batch. Finally, regarding the grid choice: Inspired by the grid choice in Han et al. (2020), we adopt a similar but somewhat different static grid for our setting:

$$t_1 = b\sqrt{s_0}, \qquad t_m = \lfloor b\sqrt{t_{m-1}}\rfloor, \qquad m\in\{2,3,\dots,M\},$$

where $b = \Theta(\sqrt{T}\cdot(T/s_0)^{\frac{1}{2(2^M-1)}})$ is chosen such that $t_M = T$. The complete algorithm is described in Algorithm 1. We emphasize again this static grid choice, rather than a dynamic one, is not a limitation of our algorithm: As we

discuss next, it is sufficient to achieve the optimal regret bound (up to log factors) for the class of dynamic batch learning algorithms.

Algorithm 1 (LBGL Under Model-C)

Input Time horizon T; context dimension d; number of batches M; sparsity bound s_0.

Initialize $b = \Theta\left(\sqrt{T}\cdot(T/s_0)^{\frac{1}{2(2^M-1)}}\right)$; $\hat{\theta}_0 = \mathbf{0} \in \mathbb{R}^d$;

Static grid $\mathcal{T} = \{t_1, \ldots, t_M\}$, with $t_1 = b\sqrt{s_0}$ and $t_m = b\sqrt{t_{m-1}}$ for $t \in \{2, \ldots, M\}$;

Partition each batch into M intervals evenly, that is, $(t_{m-1}, t_m] = \cup_{j=1}^{M} T_m^{(j)}$, for $m \in [M]$.

for $m \leftarrow 1$ ***to*** M **do**

  for $t \leftarrow t_{m-1}$ ***to*** t_m **do**

    (a) Choose $a_t = \arg\max_{a\in[K]} x_{t,a}^\top \hat{\theta}_{m-1}$ (break ties with lower action index).

    (b) Incur reward r_{t,a_t}.

  end

  $T^{(m)} \leftarrow \cup_{m'=1}^{m} T_{m'}^{(m)}$;

  $\lambda_m \leftarrow 10\sqrt{\dfrac{2\log K(\log d + 2\log T)}{|T^{(m)}|}}$;

  Update $\hat{\theta}_m \leftarrow \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}} (r_{t,a_t} - x_{t,a_t}^\top \theta)^2 + \lambda_m\|\theta\|_1$.

end

Theorem 2 characterizes the performance of the LBGL algorithm. In this section, we present the main steps in proving Theorem 2, leaving the details to Appendix D.

Theorem 2. *Under* Model-C, *Assumptions 1–4 and* $M = O(\text{loglog}(T/s_0))$, *we have*

$$\sup_{\theta^*:\|\theta^*\|_2\le 1, \|\theta^*\|_0\le s_0} \mathbb{E}_{\theta^*}[R_T(\mathbf{Alg})] \le \frac{C\cdot M^{3/2}\sqrt{\log K \log(KT)\log(dT)}}{\gamma(K)\rho(K)}\cdot\sqrt{Ts_0}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}, \tag{10}$$

where ***Alg*** *is LBGL and* $C > 0$ *is a numerical constant independent of* (T, d, M, K, s_0).

Remark 8. This regret upper bound matches the lower bound proved in Theorem 1 (up to logarithmic factors). That we only stated the theorem for $M = O(\text{loglog}(T/s_0))$ is not a restriction, but instead a merit of our result: With the number of batches $M = O(\text{loglog}(T/s_0))$, we are already able to achieve the fully online optimal regret (up to log factors) $\tilde{O}(\sqrt{Ts_0})$. Lemma 4 established the $\Omega(\sqrt{Ts_0})$ lower bound for fully online learning (under $K = 2$) and hence a matching $\tilde{O}(\sqrt{Ts_0})$ regret bound indicates that it is minimax optimal. Consequently, for any larger M, the achievable regret, which a priori will not get worse, cannot get better.

The regret of any dynamic batch learning algorithm can be achieved by a fully online learning algorithm—in the online setting you can always divide the observations into batches and run the corresponding batch algorithm—and this observation immediately yields Corollary 1.

Corollary 1. *In the fully online learning setting* $(M = T)$ *and under Assumptions 1–4:*

$$\sup_{\theta^*:\|\theta^*\|_2\le 1, \|\theta^*\|_0\le s_0} \mathbb{E}_{\theta^*}[R_T(\mathbf{Alg})] \le \frac{C\sqrt{\left(\log\log(T/s_0)\right)^3 \log K \log(KT)\log(dT)}}{\gamma(K)\rho(K)}\cdot\sqrt{Ts_0}, \tag{11}$$

where $C > 0$ *is a numerical constant independent of* (T, d, M, K, s_0).

4.2. Regret Analysis

In this section, we present the main steps of proving Theorem 2. We start by showing that the empirical covariance matrices are well conditioned even when the arms are adaptively chosen: In particular, although unlike in the low-dimensional settings the empirical covariance matrices are rank-deficient (as a result of high-dimensional features), the restricted eigenvalues are well behaved. Then we leverage standard Lasso results to show that with "well-behaved" empirical covariance matrices, the Lasso estimates of θ^* is reasonably close to the true parameters. Finally we translate the above results into the regret analysis, and establish the desired regret upper bound.

4.2.1. Establishing the Restricted Eigenvalue Condition. Given a sparsity parameter s and a matrix A, we define the key quantity *restricted eigenvalues*:

$$\phi_{\min}(s, A) \triangleq \min_{v\in\mathbb{R}^d:\|v\|_0\le s}\left\{\frac{v^\top A v}{\|v\|_2^2}\right\},$$

$$\phi_{\max}(s, A) \triangleq \max_{v\in\mathbb{R}^d:\|v\|_0\le s}\left\{\frac{v^\top A v}{\|v\|_2^2}\right\}.$$

Following the notation in Algorithm 1, $T_m^{(j)}$ denotes the jth interval of the mth batch (where the mth batch has been divided evenly into M intervals). We then define for any $j, m \in [M]$ the empirical covariance matrix: $D_{m,j} = \sum_{t\in T_m^{(j)}} x_{t,a_t} x_{t,a_t}^\top$ and $A_m = \sum_{j=1}^{m} D_{j,m}$. Lemma 5 shows that the restricted eigenvalues are bounded from both above and below with high probabilities.

Lemma 5. *Suppose Assumptions 1–4 hold. Given a sparsity parameter* s, *with probability at least* $1 - 2M^2\exp(O(s\log d) - \Omega(\rho^2(K)\cdot\sqrt{Ts_0}/M))$, *for any* $j, m \in [M]$,

$$\phi_{\max}\left(s, \frac{D_{m,j}}{|T_m^{(j)}|}\right) \le 16\log K,$$

$$\phi_{\min}\left(s, \frac{D_{m,j}}{|T_m^{(j)}|}\right) \ge \frac{\gamma(K)\rho(K)}{4}.$$

The detailed proof of Lemma 5 is deferred to Appendix D.1, and we provide the high-level steps here. For a given $v \in \mathbb{R}^d$ such that $\|v\|_0 \le s$ and $\|v\|_2 = 1$, we prove the upper bound of $v^\top D_{m,j} v$ using standard concentration inequalities. We then generalize the upper bound to an ε-net of the set of all s-sparse v by taking a union bound. Finally, we extend the result to any s-sparse v by utilizing the property of the ε-net. The proof of the lower bound is similar to that of the upper bound, except that we apply Assumption 2 when proving the lower bound for a single vector.

4.2.2. Bounding Lasso Estimation Error. With well-behaved restricted eigenvalues, Lemma 6 leverages standard Lasso results to prove an estimation error bound for $\|\hat{\theta}_m - \theta^\star\|_2$.

Lemma 6. *Under Assumptions 1–4, with probability at least* $1 - M\exp(\log d - \log K \cdot \Omega(\sqrt{Ts_0}/M)) - 2M^2 \cdot \exp\left(O\left(s_0 \cdot \frac{\log K \log d}{\gamma(K)\rho(K)}\right) - \Omega(\rho^2(K)\sqrt{Ts_0})\right) - M \cdot T^{-2}$, *for any* $m \in [M]$,

$$\|\hat{\theta}_m - \theta^\star\|_2 \le \frac{800\sqrt{2}}{\gamma(K)\rho(K)} \cdot \sqrt{s_0 M} \cdot \sqrt{\frac{\log K \cdot (2\log T + \log d)}{t_m}}.$$

The proof uses classical Lasso theory (Bickel et al. 2009) and is given in Appendix D.2.

4.2.3. Analyzing Regret Upper Bound. With Lemmas 5 and 6, we are now ready to bound the regret of Algorithm 1. Given $m \in [M]$, consider $t \in \{t_{m-1}+1, \ldots, t_m\}$, the instantaneous regret can be bounded as $\max_{a\in[K]} (x_{t,a} - x_{t,a_t})^\top \theta^\star \le \max_{a\in[K]} (x_{t,a} - x_{t,a_t})^\top (\theta^\star - \hat{\theta}_{m-1}) \le 2\max_{a\in[K]} |x_{t,a}^\top (\theta^\star - \hat{\theta}_{m-1})|$, where the first inequality is from the definition of a_t.

For a fixed $a \in [K]$, $x_{t,a}^\top (\theta^\star - \hat{\theta}_{m-1})$ is $\|\theta^\star - \hat{\theta}_{m-1}\|_2^2$-sub-Gaussian. Thus, applying a sub-Gaussian maximal inequality, we get that given a $t \in [T]$, with probability at least $1 - T^{-3}$:

$$2\max_{a\in[K]} |x_{t,a}^\top (\theta^\star - \hat{\theta}_{m-1})| \le 6\sqrt{\log(TK)} \cdot \|\theta^\star - \hat{\theta}_{m-1}\|_2.$$

Applying a union bound over the batch m with $m \ge 2$ and invoking Lemma 6, we have with probability at least $1 - (1+M)\cdot T^{-2} - M \cdot \exp(\log d - \log K \cdot \Omega(\sqrt{Ts_0}/M)) - 2M^2 \cdot \exp\left(O\left(s_0 \frac{\log K \log d}{\gamma(K)\rho(K)}\right) - \Omega(\rho^2(K) \cdot \sqrt{Ts_0}/M)\right)$,

$$\max_{a\in[K]} (x_{t,a} - x_{t,a_t})^\top \theta^\star \le \frac{C}{\gamma(K)\rho(K)} \cdot \sqrt{s_0 M \log(TK)} \cdot \sqrt{\frac{\log K(2\log T + \log d)}{t_{m-1}}}, \quad \forall t \in [t_{m-1}+1, t_m],$$

where $C > 0$ is a numerical constant. Summing over the regret incurred in the $m \ge 2$ batches yields

$$\begin{aligned}
&\sum_{m=2}^{M} \sum_{t=t_{m-1}+1}^{t_m} \max_{a\in[K]} (x_{t,a} - x_{t,a_t})^\top \theta^\star \\
&\le \frac{C}{\gamma(K)\rho(K)} \cdot bM^{3/2} \cdot \sqrt{s_0 \log K \log(TK)(\log d + 2\log T)} \\
&\le \frac{C'}{\gamma(K)\rho(K)} \cdot M^{3/2} \\
&\quad \cdot \sqrt{\log K \log(TK)(\log d + 2\log T)} \sqrt{Ts_0} \left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},
\end{aligned}$$

where $b = \Theta\left(\sqrt{T} \cdot (T/s_0)^{\frac{1}{2(2^M-1)}}\right)$ is from the choice of grids. Finally for the first batch, because no rewards are observed, it suffices for us to adopt a crude bound:

$$\sum_{t=1}^{t_1} \max_{a\in[K]} (x_{t,a} - x_{t,a_t})^\top \theta^\star \le 2\sum_{t=1}^{t_1} \left(\max_{a\in[K]} x_{t,a}^\top \theta^\star\right).$$

Applying a sub-Gaussian maximal inequality and a union bound over all $t \in [t_1]$, we have with probability at least $1 - T^{-2}$,

$$\begin{aligned}
\sum_{t=1}^{t_1} \max_{a\in[K]} (x_{t,a} - x_{t,a_t})^\top \theta^\star &\le 6\sqrt{\log(KT)} \cdot t_1 \\
&= \Theta\left(\sqrt{\log(KT)} \sqrt{Ts_0} \cdot \left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}\right).
\end{aligned}$$

Putting everything together, we then have that with probability at least $1 - (2+M)\cdot T^{-2} - 2M^2 \cdot \exp\left(O\left(s_0 \frac{\log K \log d}{\gamma(K)\rho(K)}\right) - \Omega(\rho^2(K) \cdot \sqrt{Ts_0}/M)\right) - M \cdot \exp(\log d - \log K \cdot \Omega(\sqrt{Ts_0}/M))$,

$$\begin{aligned}
R_T(\mathbf{Alg}) &\le \frac{C''}{\gamma(K)\rho(K)} \cdot M^{3/2} \cdot \sqrt{\log K \log(KT) \log(dT)} \\
&\quad \cdot \sqrt{Ts_0} \cdot \left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},
\end{aligned}$$

where $C'' > 0$ is a numerical constant resulting from merging the constant corresponding to the first batch and the constant C' (corresponding to all subsequent batches). Because $M = O(\log\log(T/s_0))$, the previous high-probability regret bound immediately implies the expected regret bound:

$$\begin{aligned}
\mathbb{E}_{\theta^\star}[R_T(\mathbf{Alg})] &\le \frac{C'''}{\gamma(K)\rho(K)} \cdot M^{3/2} \cdot \sqrt{\log K \log(KT) \log(dT)} \\
&\quad \cdot \sqrt{Ts_0} \cdot \left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},
\end{aligned}$$

where $C''' > 0$ is a numerical constant.

5. Discussion

Through matching lower and upper regret bounds, our work completes (up to certain log factors) the picture of

dynamic batch learning in high-dimensional sparse linear contextual bandits. Furthermore, the algorithm provided is very simple to implement in practice, an important merit from a practical standpoint. We close the paper by discussing possible extensions of our work.

5.1. Extension to Subexponential Reward Distribution

In this paper, we focused on sub-Gaussian reward distribution. It would be interesting to consider the high-dimensional dynamic batch learning problem with subexponential reward distribution (although this is a hard task even in the fully online setting).

5.2. Extension to Sparsity-Agnostic Algorithm

It would be desirable to have a dynamic batch learning algorithm that would not require any knowledge of a sparsity upper bound. In the fully online decision-making setting, Oh et al. (2021) propose such an algorithm. Adapting it the batched setting, however, is challenging because the grid design critically depends on s_0.

5.3. Other Extensions

It would be interesting to explore the continuous action set case and understand whether learning guarantees in this regime are materially worse. Finally, going beyond to the nonparametric contextual bandits setting would also be useful.

Appendix A. Definitions and Auxiliary Results

We collect in this section all the known results in the existing literature that will be useful for us.

Definition A.1. Let $(\mathcal{X},\mathcal{F})$ be a measurable space and P, Q be two probability measures on $(\mathcal{X},\mathcal{F})$.

(a) The total-variation distance between P and Q is defined as

$$\mathrm{TV}(P,Q)=\sup_{A\in\mathcal{A}}|P(A)-Q(A)|.$$

(b) The KL divergence between P and Q is

$$D_{\mathrm{KL}}(P\|Q)=\begin{cases}\int\log\frac{dP}{dQ}dP & \text{if } P\ll Q\\ +\infty & \text{otherwise}\end{cases}.$$

Lemma A.1 (Paley-Zygmund Inequality). *If $X\ge 0$ is a random variable whose variance is finite, Then for any $\theta\in(0,1)$,*

$$(1)\ \mathbb{P}(X>\theta\mathbb{E}[X])\ge(1-\theta)^2\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]};$$
$$(2)\ \mathbb{P}(X>\theta\mathbb{E}[X])\ge\frac{(1-\theta)^2\mathbb{E}[X]^2}{\mathrm{Var}(Z)+(1-\theta)^2\mathbb{E}[X]^2}.$$

Lemma A.2 (Data-Processing Inequality (Cover and Thomas 2006)). *Let X, Y, Z denote random variables drawn from a Markov chain in the order (denoted by $X\to Y\to Z$) that the conditional distribution of Z depends only on Y and is conditionally independent of X. Then if $X\to Y\to Z$, we have $I(X;Y)\ge I(X;Z)$, where $I(X;Y)$ is the mutual information between X and Y.*

Lemma A.3 (Pinsker's Inequality). *Let P and Q be any two probability measures on the same measurable space. Then* $\mathrm{TV}(P,Q)\le\sqrt{\frac{1}{2}\cdot D_{\mathrm{KL}}(P\|Q)}$.

Lemma A.4 (Joint Convexity of the KL divergence (Cover and Thomas 2006)). *The KL divergence $D_{\mathrm{KL}}(P||Q)$ is jointly convex in its argument P and Q: let P_1,P_2,Q_1,Q_2 be distributions on $\mathcal{X}$, then for any $\lambda\in[0,1]$,*

$$D_{\mathrm{KL}}(\lambda P_1+(1-\lambda)P_2\|\lambda Q_1+(1-\lambda)Q_2)\le\lambda D_{\mathrm{KL}}(P_1\|Q_1)+(1-\lambda)D_{\mathrm{KL}}(P_2\|Q_2).$$

Lemma A.5 (Sub-Gaussian Maximal Inequality (Rigollet 2015)). *Let $X_1,\dots,X_K$ be K centered σ^2-sub-Gaussian random variables, then for any $t>0$, $\mathbb{P}(\max_{k\in[K]}X_k\ge t)\le Ke^{-\frac{t^2}{2\sigma^2}}$.*

Appendix B. Proof of Lemma 1

1. The proof follows directly from Han et al. (2020, lemma 4).

2. Given a unit vector $v\in\mathbb{R}^d$, for any $\delta>0$,

$$\mathbb{P}\left((v^\top x_{t,a_t})^2\le\frac{\alpha}{e}\cdot(2cK)^{-\frac{1}{\alpha}}\right)=\mathbb{P}\left(-(v^\top x_{t,a_t})^2\ge-\frac{\alpha}{e}\cdot(2cK)^{-\frac{1}{\alpha}}\right)$$
$$=\mathbb{P}\left(\exp(-(v^\top x_{t,a_t})^2\cdot\delta)\ge\exp\left(-\frac{\alpha}{e}\cdot(2cK)^{-\frac{1}{\alpha}}\cdot\delta\right)\right)$$
$$\overset{(a)}{\le}\exp\left(\frac{\alpha}{e}\cdot\delta\cdot(2cK)^{-\frac{1}{\alpha}}\right)\cdot\mathbb{E}[\exp(-(v^\top x_{t,a_t})^2\cdot\delta)]$$
$$\le\exp\left(\frac{\alpha}{e}\cdot\delta\cdot(2cK)^{-\frac{1}{\alpha}}\right)\cdot\sum_{a\in[K]}\mathbb{E}[\exp(-(v^\top x_{t,a})^2\cdot\delta)]$$
$$\overset{(b)}{\le}cK\cdot\exp\left(\frac{\alpha}{e}\cdot\delta\cdot(2cK)^{-\frac{1}{\alpha}}\right)\cdot\delta^{-\alpha},$$

where step (a) is due to Markov's inequality and step (b) follows from Equation (1). Taking $\delta=e\cdot(2cK)^{1/\alpha}$ (the minimizer of the upper bound), we arrive at $\mathbb{P}((v^\top x_{t,a_t})^2\ge\frac{\alpha}{e}\cdot(2cK)^{-1/\alpha})\ge\frac{1}{2}$.

3. Let $v\in\mathbb{R}^d$ be an arbitrary unit vector. For each $a\in[K]$,

$$\mathbb{P}\left((v^\top x_{t,a})^2\ge\frac{1}{2}\cdot\mathbb{E}[(v^\top x_{t,a})^2]\right)$$
$$\overset{(a)}{\ge}\frac{\frac{1}{4}\cdot\mathbb{E}[(v^\top x_{t,a})^2]^2}{\mathrm{Var}((v^\top x_{t,a})^2)+\frac{1}{4}\cdot\mathbb{E}[(v^\top x_{t,a})^2]^2}$$
$$\overset{(b)}{\ge}\frac{\frac{1}{4}\cdot\mathbb{E}[(v^\top x_{t,a})^2]^2}{\frac{\Lambda^2}{8K}++\frac{1}{4}\cdot\mathbb{E}[(v^\top x_{t,a})^2]^2}\overset{(c)}{\ge}\frac{2K}{2K+1}.$$

Previously, step (a) is due to the Paley-Zygmund inequality (Lemma A.1); steps (b) and (c) follow from the assumption. As a consequence, we have $\mathbb{P}((v^\top x_{t,a})^2<\frac{\Lambda}{2})\le\frac{1}{2K+1}$. Finally,

$$\mathbb{P}\left((v^\top x_{t,a_t})^2\ge\frac{\Lambda}{2}\right)$$
$$\ge\mathbb{P}\left(\min_{a\in[K]}(v^\top x_{t,a})^2\ge\frac{\Lambda}{2}\right)=1-\mathbb{P}\left(\min_{a\in[K]}(v^\top x_{t,a})^2<\frac{\Lambda}{2}\right)$$
$$\ge1-\sum_{a\in[K]}\mathbb{P}\left((v^\top x_{t,a})^2<\frac{\Lambda}{2}\right)\ge1-\frac{K}{2K+1}\ge\frac{1}{2},$$

completing the proof.

4. Without loss of generality, we assume $\nu \le 1$. For an arbitrary unit vector $v \in \mathbb{R}^d$,

$$\mathbb{P}\left((v^\top x_{t,a_t})^2 \ge \frac{\Lambda}{2}\right) = \sum_{a=1}^{2} \mathbb{P}\left(a_t = a, (v^\top x_{t,a})^2 \ge \frac{\Lambda}{2}\right).$$

By symmetry, we only need to focus on $a=1$, for which we have

$$\begin{aligned}
&\mathbb{P}\left(a_t = 1, (v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right) = \int \mathbf{1}\left\{x_{t,a}^\top \theta \ge x_{t,2}^\top \theta, (v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right\} \\
&\quad \cdot p(x_{t,1}, x_{t,2}) dx_{t,1} dx_{t,2} \\
&\overset{(a)}{\ge} \frac{1}{2} \cdot \mathbb{P}\left(a_t = 1, (v^\top x_{t,a}) \ge \frac{\Lambda}{2}\right) + \frac{\nu}{2} \int \mathbf{1}\left\{x_{t,1}^\top \theta \ge x_{t,2}^\top \theta, (v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right\} \\
&\quad \cdot p(-x_{t,1}, -x_{t,2}) dx_{t,1} dx_{t,2} \\
&= \frac{1}{2} \cdot \mathbb{P}\left(a_t = 1, (v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right) + \frac{\nu}{2} \cdot \mathbb{P}\left(a_t = 2, (v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right) \\
&\ge \frac{\nu}{2} \cdot \mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right) \ge \frac{\nu}{2} \cdot \mathbb{P}\left((v^\top x_{t,a})^2 \ge \frac{\mathbb{E}[(v^\top x_{t,1})^2]}{2}\right) \\
&\overset{(b)}{\ge} \frac{\nu}{8} \cdot \frac{\mathbb{E}[(v^\top x_{t,a})^2]^2}{\mathbb{E}[(v^\top x_{t,a})^4]} \overset{(c)}{\ge} \frac{\nu \Lambda^2}{128},
\end{aligned}$$

where step (a) is by the assumption; step (a) is due to the Paley-Zygmund inequality; and step (c) is because of the assumption and $v^\top x_{t,a}$ is 1-sub-Gaussian. Combining the case of $a=1$ and $a=2$, we have

$$\mathbb{P}\left((v^\top x_{t,a_t})^2 \ge \frac{\Lambda}{2}\right) \ge \frac{\nu \Lambda^2}{64}.$$

5. Without loss of generality, we assume $\nu_1, \nu_2(K) \le 1$. Fix an arbitrary unit vector v, and $\theta \in \mathbb{R}^d$.

To start, we focus on $a=1$. We decompose all the permutations of $[K]$ into three subsets: $\mathcal{I}_{\min}, \mathcal{I}_{\max}$ and $\mathcal{I}_{\text{mid}}$, where $\mathcal{I}_{\min} := \{\pi : \pi_1 = 1\}, \mathcal{I}_{\max} := \{\pi : \pi_K = 1\}$ and $\mathcal{I}_{\text{mid}} = \{\pi : \pi_1 \ne 1, \pi_K \ne 1\}$. We then have

$$\begin{aligned}
\mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right) &= \sum_{\pi : \pi \in \mathcal{I}_{\min}} \mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right) \\
&\quad + \sum_{\pi : \pi \in \mathcal{I}_{\max}} \mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right) \\
&\quad + \sum_{\pi : \pi \in \mathcal{I}_{\text{mid}}} \mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right).
\end{aligned}$$

By the assumption, for any permutation π,

$$\begin{aligned}
&\nu_2(K) \cdot \mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right) \\
&\le \mathbb{P}\left((v^\top x_{t,\pi_1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right) \\
&\quad + \mathbb{P}\left((v^\top x_{t,\pi_K})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right).
\end{aligned}$$

As a consequence,

$$\begin{aligned}
&\mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right) \le \frac{1}{\nu_2(K)} \\
&\quad \cdot \sum_{\pi} \mathbb{P}\left((v^\top x_{t,\pi_1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right) \\
&\quad + \mathbb{P}\left((v^\top x_{t,\pi_K})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right).
\end{aligned}$$

Previously, by the relaxed symmetry condition,

$$\begin{aligned}
&\mathbb{P}\left((v^\top x_{t,\pi_1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right) \\
&\le \frac{1}{\nu_1} \cdot \int \mathbf{1}\left\{(v^\top x_{t,\pi_1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right\} \\
&\quad \cdot p(-x_{t,1}, \ldots, -x_{t,K}) dx_{t,1} \ldots dx_{t,K} \\
&= \frac{1}{\nu_1} \cdot \mathbb{P}\left((v^\top x_{t,\pi_1})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \ge \cdots \ge x_{t,\pi_K}^\top \theta\right).
\end{aligned}$$

We then have

$$\begin{aligned}
&\mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right) \\
&\le \frac{1}{\nu_1 \nu_2(K)} \cdot \sum_{\pi} \left(\mathbb{P}\left((v^\top x_{t,a_t})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \ge \cdots \ge x_{t,\pi_K}^\top \theta\right)\right. \\
&\quad \left. + \mathbb{P}\left((v^\top x_{t,a_t})^2 \ge \frac{\Lambda}{2}, x_{t,\pi_1}^\top \theta \le \cdots \le x_{t,\pi_K}^\top \theta\right)\right) \\
&= \frac{2}{\nu_1 \nu_2(K)} \cdot \mathbb{P}\left((v^\top x_{t,a_t})^2 \ge \frac{\Lambda}{2}\right).
\end{aligned}$$

Finally, we arrive at

$$\begin{aligned}
\mathbb{P}\left((v^\top x_{t,a_t})^2 \ge \frac{\Lambda}{2}\right) &\ge \frac{\nu_1 \nu_2(K)}{2} \cdot \mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{\Lambda}{2}\right) \\
&\ge \frac{\nu_1 \nu_2(K)}{2} \cdot \mathbb{P}\left((v^\top x_{t,1})^2 \ge \frac{v^\top \mathbb{E}[x_{t,1} x_{t,1}^\top] v}{2}\right) \\
&\overset{(a)}{\ge} \frac{\nu_1 \nu_2(K)}{8} \frac{\Lambda^2}{16} = \frac{\nu_1 \nu_2(K) \Lambda^2}{128},
\end{aligned}$$

where step (a) is due to the Paley-Zygmund (Lemma A.1) inequality and the assumption.

Appendix C. Proof of Main Lemmas in Section 3

C.1. Proof of Lemma 4

As in the batched case (and with the same notation), we construct a prior Q for θ, where $\theta(S) \sim \mathsf{Unif}(\Delta \mathbb{S}^{s_0 - 1})$ and $\theta^\star(S^c) = 0$, and $\Delta = \sqrt{s_0/32T}$. Then,

$$\begin{aligned}
&\sup_{\theta^\star : \|\theta^\star\|_2 \le 1, \|\theta^\star\|_0 \le s_0} \mathbb{E}_{\theta^\star}[R_T(\mathbf{Alg})] \\
&\ge \mathbb{E}_Q \mathbb{E}_\theta[R_T(\mathbf{Alg})] = \sum_{t=1}^{T} \mathbb{E}_Q \mathbb{E}_x \mathbb{E}_{P_{\theta,x}^t}\left[\max_{a \in \{1,2\}} (x_{t,a}^\top \theta - x_{\tau,a_\tau}^\top \theta)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_x \mathbb{E}_Q \mathbb{E}_{P_{\theta,x}^t}[\mathbf{1}\{a_t = 1\} \cdot (d_t^\top \theta)_+ + \mathbf{1}\{a_t = 2\} \cdot (d_t^\top \theta)_-],
\end{aligned} \tag{C.1}$$

where $d_t = x_{t,2} - x_{t,1}$. Define two new measures via: $\frac{dQ_t^+}{dQ}(\theta) = \frac{(d_t^\top \theta)_+}{Z(d_t)}$ and $\frac{dQ_t^-}{dQ}(\theta) = \frac{(d_t^\top \theta)_-}{Z(d_t)}$; $Z(d_t) = \mathbb{E}_Q[(d_t^\top \theta)_+] = \mathbb{E}_Q[(d_t^\top \theta)_-]$ is a common normalizing constant. Using this representation, we

have

$$\begin{aligned}(\text{C.1}) &= \sum_{t=1}^{T}\mathbb{E}_x[Z(d_t)\cdot(\mathbb{E}_{P_{\theta_0,x}\circ Q_t^+}[\mathbf{1}\{a_t=1\}]+\mathbb{E}_{P_{\theta_0,x}\circ Q_t^-}[\mathbf{1}\{a_t=2\}])]\\ &\overset{(a)}{\geq}\sum_{t=1}^{T}\mathbb{E}_x[Z(d_t)\cdot(1-\mathsf{TV}(P_{\theta,x}^{t-1}\circ Q_t^+,P_{\theta,x}^{t-1}\circ Q_t^-))]\\ &\overset{(b)}{\geq}\sum_{t=1}^{T}\mathbb{E}_x\left[Z(d_t)\cdot\left(1-\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(P_{\theta,x}^{t-1}\circ Q_t^+\,\|\,P_{\theta-2(\tilde{u}_t^\top\theta)\tilde{u}_t,x}^{t-1}\circ Q_t^+\right)}\right)\right]\\ &\overset{(c)}{\geq}\sum_{t=1}^{T}\mathbb{E}_x\left[Z(d_t)\cdot\left(1-\sqrt{\frac{1}{2}\mathbb{E}_{Q_t^+}[D_{\mathrm{KL}}\left(P_{\theta,x}^{t-1}\,\|\,P_{\theta-2(\tilde{u}_t^\top\theta)\tilde{u}_t,x}^{t-1}\right)]}\right)\right],\end{aligned} \tag{C.2}$$

where step (a) follows from $P(A)+Q(A^c)\leq 1-\mathsf{TV}(P,Q)$; step (b) is due to a change of measure and Lemma A.3; and step (c) is because of the joint convexity of the KL divergence. Previously,

$$\begin{aligned}D_{\mathrm{KL}}\left(P_{\theta,x}^{t-1}\|P_{\theta-2(\tilde{u}_t^\top\theta)\tilde{u}_t,x}^{t-1}\right)&=\frac{1}{2}\sum_{\tau=1}^{t-1}\left(2(\tilde{u}_t^\top\theta)\cdot(\tilde{u}_t^\top x_{\tau,a_\tau})\right)^2\\ &=2(\tilde{u}_t^\top\theta)^2\cdot\tilde{u}_t^\top\left(\sum_{\tau=1}^{t-1}x_{\tau,a_\tau}x_{\tau,a_\tau}^\top\right)\tilde{u}_t,\end{aligned}$$

where $\tilde{u}_t\in\mathbb{R}^d$ satisfies $\tilde{u}_t(S)=\frac{d_t(S)}{\|d_t(S)\|_2}$ and $\tilde{u}_t(S^c)=0$. Plugging in the expression of the KL divergence, we have

$$\begin{aligned}(\text{C.2})&\geq\sum_{t=1}^{T}\mathbb{E}_x\left[Z(d_t)\cdot\left(1-\sqrt{\mathbb{E}_{Q_t^+}[(\tilde{u}_t^\top\theta)^2]\cdot\tilde{u}_t^\top\left(\sum_{\tau=1}^{t-1}x_{\tau,a_\tau}^\top x_{\tau,a_\tau}\right)\tilde{u}_t}\right)\right]\\ &\geq\sum_{t=1}^{T}\mathbb{E}_x\left[Z(d_t)\cdot\left(1-\sqrt{\mathbb{E}_{Q_t^+}[(\tilde{u}_t^\top\theta)^2]\cdot\tilde{u}_t^\top\left(\sum_{\tau=1}^{t-1}x_{\tau,1}x_{\tau,1}^\top+x_{\tau,2}x_{\tau,2}^\top\right)\tilde{u}_t}\right)\right]\\ &\overset{(a)}{\geq}\sum_{t=1}^{T}\mathbb{E}_x\left[Z(d_t)\cdot\left(1-\sqrt{\frac{4t\Delta^2}{s_0}}\right)\right]\overset{(b)}{\geq}\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_x[Z(d_t)]\geq\frac{T\Delta}{10}=\frac{\sqrt{Ts_0}}{40\sqrt{2}},\end{aligned}$$

where step (a) is by taking expectation w.r.t. $\{(x_{\tau,1},x_{\tau,2})\}_{\tau\leq t-1}$; and step (b) follows from the choice of Δ. The proof is completed.

Appendix D. Proof of Main Lemmas in Section 4

D.1. Proof of Lemma 5

Consider the mth batch, for any $j\in[M]$, by definition $a_t=\arg\max_{a\in[K]}x_{t,a}^\top\hat{\theta}_{m-1}$ for any $t\in T_m^{(j)}$, where $\hat{\theta}_{m-1}$ depends *only* on the observations from batch 1 to $m-1$. Hence $\{x_{t,a_t}\}_{t\in T_m^{(j)}}$ are mutually independent and follow the same distribution conditional on the previous batches. Consider now a fixed sparsity upper bound s.

D.1.1. Upper Bound. Given a vector $v\in\mathbb{R}^d$, such that $\|v\|_0\leq s$ and $\|v\|_2=1$. Let $\mathsf{supp}(v)$ denote the support of v, where without loss of generality we assume $|\mathsf{supp}(v)|=s$ (otherwise we can include extra zero coordinates in $\mathsf{supp}(v)$), and let $\mathcal{N}(\varepsilon)$ denote the ε-net of $\mathbb{S}^{s-1}$. For notational simplicity, denote $Y_{t,a}=(v^\top x_{t,a})^2$. For any $\delta,\mu>0$, one has

$$\begin{aligned}&\mathbb{P}\left(\sum_{t\in T_m^{(j)}}Y_{t,a_t}\geq(4+\delta)\cdot|T_m^{(j)}|\,|\,\hat{\theta}_{m-1}\right)\\ &\overset{(a)}{\leq}\exp\left(-\mu(4+\delta)\cdot|T_m^{(j)}|\right)\cdot\mathbb{E}\left[\exp\left(\mu\cdot\sum_{t\in T_m^{(j)}}Y_{t,a_t}\right)|\,\hat{\theta}_{m-1}\right]\\ &\overset{(b)}{=}\exp\left(-\mu(4+\delta)\cdot|T_m^{(j)}|\right)\cdot\prod_{t\in T_m^{(j)}}\mathbb{E}\left[\exp(\mu\cdot Y_{t,a_t})\,|\,\hat{\theta}_{m-1}\right]\\ &\leq\exp\left(-\mu(4+\delta)\cdot|T_m^{(j)}|\right)\prod_{t\in T_m^{(j)}}\left(\sum_{a\in[K]}\mathbb{E}[\exp(\mu Y_{t,a})]\right)\end{aligned}$$

where step (a) follows from the Markov's inequality, and step (b) is due to the (conditional) independence across t. Because $x_{t,a}$ is 1-sub-Gaussian, $v^\top x_{x,a}$ is as well 1-sub-Gaussian. As a result, $Y_{t,a}-\mathbb{E}[Y_{t,a}]$ is subexponential with parameter $(4\sqrt{2},4)$, and $\mathbb{E}[Y_{t,a}]\leq 4$. With this, we obtain a Bernstein-type bound:

$$\begin{aligned}&\mathbb{P}\left(\sum_{t\in T_m^{(j)}}Y_{t,a_t}\geq(4+\delta)\cdot|T_m^{(j)}|\right)\\ &\leq\exp\left(\left(-\min\left(\frac{\delta^2}{64},\frac{\delta}{8}\right)+\log K\right)\cdot|T_m^{(j)}|\right).\end{aligned}$$

Combining everything previously and letting $\delta=9\log K$, we arrive at

$$\mathbb{P}\left(\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}(v^\top x_{t,a_t})^2\geq 4+9\log K\right)\leq\exp\left(-\frac{\log K}{8}\cdot|T_m^{(j)}|\right).$$

Taking a union bound, we get that with probability at least $1-\exp(s\log d+s\log(1+1/\varepsilon)-|T_m^{(j)}|\log K/8)$, for any v such that $\|v\|_0\leq s,\|v\|_2=1$ and $v(\mathsf{supp}(v))\in\mathcal{N}(\varepsilon)$,

$$\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}(v^\top x_{t,a_t})^2\leq 4+9\log K\leq 15\log K.$$

Let $u\in\mathbb{R}^d$ be an *arbitrary* vector such that $\|u\|_0\leq s$ and $\|u\|_2=1$. By the definition of the ε-net, there exists $v_0\in\mathcal{N}(\varepsilon)$, such that $\|u(\mathsf{supp}(u))-v_0\|_2\leq\varepsilon$. Let $v\in\mathbb{R}^d$ be a vector such that $v(\mathsf{supp}(u))=v_0$ and $v(\mathsf{supp}(u)^c)=0$. By construction $\|u-v\|_2\leq\varepsilon$. Consequently,

$$\begin{aligned}\frac{u^\top D_{m,j}u}{|T_m^{(j)}|}-\frac{v^\top D_{m,j}v}{|T_m^{(j)}|}&=\frac{u^\top D_{m,j}(u-v)}{|T_m^{(j)}|}+\frac{(u-v)^\top D_{m,j}v}{|T_m^{(j)}|}\\ &\leq 2\varepsilon\phi_{\max}\left(s,\frac{D_{m,j}}{|T_m^{(j)}|}\right).\end{aligned}$$

Note that $|T_m^{(j)}|=\Omega(\sqrt{Ts_0}/M)$, for any $j,m\in[M]$. Taking the supreme over u and rearranging yields that with probability at least $1-\exp(s\log d+s\log(1+1/\varepsilon)-\Omega(\sqrt{Ts_0}/M))$,

$$\phi_{\max}\left(s,\frac{D_m}{|T_m^{(j)}|}\right)\leq\frac{15\log K}{1-2\varepsilon}. \tag{D.1}$$

D.1.2. Lower Bound. We now proceed to prove a lower bound for the restricted eigenvalues. By Assumption 2,

$\mathbb{P}(Y_{t,a_t} \geq \gamma(K)|\hat{\theta}_{m-1}) \geq \rho(K)$. We then have that

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}} Y_{t,a_t} \leq \frac{\gamma(K)\rho(K)}{2} \mid \hat{\theta}_{m-1}\right)\\
&\leq \mathbb{P}\left(\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}} \mathbf{1}\{Y_{t,a_t} \geq \gamma(K)\} \leq \frac{\rho(K)}{2} \mid \hat{\theta}_{m-1}\right)\\
&\leq \mathbb{P}\left(\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}} \mathbf{1}\{Y_{t,a_t} \geq \gamma(K)\} - \mathbb{P}(Y_{t,a_t} \geq \gamma(K)|\hat{\theta}_{m-1}) \right.\\
&\left. \leq -\frac{\rho(K)}{2} \mid \hat{\theta}_{m-1}\right) \leq \exp\left(-\frac{\rho^2(K)}{2}\cdot|T_m^{(j)}|\right),
\end{aligned}$$

where the last inequality is due to the Chernoff bound.

Taking a union bound over all s-sparse unit vector v whose support is in $\mathcal{N}(\varepsilon)$, we conclude that with probability at least $1-\exp(s\log d + s\log(1+1/\varepsilon) - \rho^2(K)|T_m^{(j)}|/2)$, for any v whose support is in $\mathcal{N}(\varepsilon)$,

$$\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}} (v^\top x_{t,a_t})^2 > \frac{\gamma(K)\rho(K)}{2}. \tag{D.2}$$

We now condition on Events (D.1) and (D.2) and turn our attention to an arbitrary vector $u \in \mathbb{R}^d$ such that $\|u\|_0 \leq s$ and $\|u\|_2 = 1$. By the definition of ε-nets, there exists $v_0 \in \mathcal{N}(\varepsilon)$ such that $\|u(\mathsf{supp}(u)) - v_0\|_2 \leq \varepsilon$. Let $v \in \mathbb{R}^d$ be the vector such that $v(\mathsf{supp}(u)) = v_0$ and $v(\mathsf{supp}(u)^c) = 0$. Then

$$\begin{aligned}
\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}} (u^\top x_{t,a_t})^2 &\geq \frac{1}{|T_m^{(j)}|}\left(\sum_{t\in T_m^{(j)}} (v^\top x_{t,a_t})^2 + 2(u-v)x_{t,a_t}x_{t,a_t}^\top v\right)\\
&\geq \frac{\gamma(K)\rho(K)}{2} - 2\varepsilon\phi_{\max}\left(s, \frac{D_{m,j}}{|T_m^{(j)}|}\right)\\
&\geq \frac{\gamma(K)\rho(K)}{2} - \frac{30\varepsilon\log K}{1-2\varepsilon}.
\end{aligned}$$

Finally letting $\varepsilon = \min(\frac{1}{32}, \frac{\gamma(K)\rho(K)}{128\log K})$ and taking a union bound over $j,m \in [M]$, we conclude that with probability at least $1-2M^2\exp(s\log d + s\log(1+\frac{128\log K}{\gamma(K)\rho(K)}) - \Omega(\rho^2(K)\cdot\sqrt{Ts_0}/M))$, for any $j,m \in [M]$,

$$\begin{aligned}
\phi_{\min}\left(s, \frac{D_m}{|T_m^{(j)}|}\right) &\geq \frac{\gamma(K)\rho(K)}{4},\\
\phi_{\max}\left(s, \frac{D_m}{|T_m^{(j)}|}\right) &\leq 16\log K. \quad \square
\end{aligned} \tag{D.3}$$

D.2. Proof of Lemma 6

To start, we work on an upper bound on the magnitude of $x_{t,a_t,l}$ (the lth coodinate of x_{t,a_t}). Given any $m \in [M]$ and $l \in [d]$, define $M_{m,l} = \sqrt{\frac{1}{T^{(m)}}\sum_{t\in T^{(m)}} x_{t,a_t,l}^2}$. For any $\delta > 0$ and $0 < \mu < 1/4$,

$$\begin{aligned}
&\mathbb{P}(M_{m,l}^2 \geq 16\log K + \delta)\\
&\overset{(a)}{\leq} \mathbb{P}\left(\sum_{t\in T^{(m)}}\left(\max_{a\in[K]} x_{t,a,l}^2 - \mathbb{E}\left[\max_{a\in[K]} x_{t,a,l}^2\right]\right) \geq |T^{(m)}|\cdot\delta\right)\\
&\overset{(b)}{\leq} \mathbb{E}\left[\exp\left(\mu\sum_{t\in T^{(m)}}\left(\max_{a\in[K]} x_{t,a,l}^2 - \mathbb{E}\left[\max_{a\in[K]} x_{t,a,l}^2\right]\right)\right)\right]\exp(-\mu\delta|T^{(m)}|)\\
&\overset{(c)}{\leq} \prod_{t\in T^{(m)}}\left\{\sum_{a\in[K]} \mathbb{E}[\exp(\mu(x_{t,a,l}^2 - \mathbb{E}[x_{t,a,l}^2]))]\right\}\exp(-\mu\delta|T^{(m)}|)\\
&\overset{(d)}{\leq} \exp((\log K + 16\mu^2 - \mu\delta)\cdot|T^{(m)}|),
\end{aligned} \tag{D.4}$$

where step (a) is because $\mathbb{E}[\max_{a\in[K]} x_{t,a,l}^2] \leq 16\log K$; step (b) follows from the Markov's inequality; step (c) is due to the independence of $x_{t,a,l}$ across t; and step (d) is because $x_{t,a,l}^2 - \mathbb{E}[x_{t,a,l}^2]$ is $(4\sqrt{2},4)$-sub-exponential. Optimizing the right-hand side of (D.4) over $0 < \mu \leq 1/4$ and taking a union bound over $l \in [d]$, we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\max_{l\in[d]} M_{m,l}^2 \geq 16\log K + \delta\right)\\
&\leq d\exp\left(\left(\log K - \min\left(\frac{\delta^2}{64}, \frac{\delta}{8}\right)\right)\cdot|T^{(m)}|\right).
\end{aligned}$$

Taking $\delta = 9\log K$, one has that with probability at least $1-\exp(\log d - \log K\cdot|T^{(m)}|/8)$, for all $l \in [d]$.

$$M_{m,l}^2 \leq 25\log K. \tag{D.5}$$

For any $m \in [M]$, any $s \leq d$ and any $v \in \mathbb{R}^d$ such that $\|v\|_0 \leq s$ and $\|v\|_2 = 1$,

$$\frac{1}{|T^{(m)}|}v^\top A_m v = \sum_{j=1}^m \frac{|T_j^{(m)}|}{|T^{(m)}|}\left(\frac{v^\top D_{j,m} v}{|T_j^{(m)}|}\right),$$

and consequently,

$$\begin{aligned}
\phi_{\max}\left(s, \frac{A_m}{|T^{(m)}|}\right) &\leq \max_{j\in[m]} \phi_{\max}\left(s, \frac{D_{j,m}}{T_j^{(m)}}\right),\\
\phi_{\min}\left(s, \frac{A_m}{|T^{(m)}|}\right) &\geq \min_{j\in[m]} \phi_{\min}\left(s, \frac{D_{j,m}}{|T_j^{(m)}|}\right).
\end{aligned}$$

By Lemma 5, with probability at least $1-2M^2\cdot\exp\left(O(s\log\left(\frac{d\log K}{\gamma(K)\rho(K)}\right)) - \Omega(\rho^2(K)\cdot\sqrt{Ts_0}/M)\right)$, for any $j,m \in [M]$,

$$\phi_{\max}\left(s, \frac{A_m}{|T^{(m)}|}\right) \leq 16\log K, \qquad \phi_{\min}\left(s, \frac{A_m}{|T^{(m)}|}\right) \geq \frac{\gamma(K)\rho(K)}{4}. \tag{D.6}$$

By the definition of $\hat{\theta}_m$,

$$\begin{aligned}
&\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}} (r_{t,a_t} - x_{t,a_t}^\top\hat{\theta}_m)^2 + \lambda_m\|\hat{\theta}_m\|_1\\
&\leq \frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}} (r_{t,a_t} - x_{t,a_t}^\top\theta^\star)^2 + \lambda_m\|\theta^\star\|_1.
\end{aligned}$$

Rearranging yields

$$\begin{aligned}
&\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}} (x_{t,a_t}^\top\theta^\star - x_{t,a_t}^\top\hat{\theta}_m)^2 + \lambda_m\|\hat{\theta}_m\|_1\\
&\leq \lambda_m\|\theta^\star\|_1 + \frac{1}{|T^m|}\sum_{t\in T^{(m)}} (x_{t,a_t}^\top\hat{\theta}_m - x_{t,a_t}^\top\theta^\star)\varepsilon_t.
\end{aligned}$$

By the construction of $T^{(m)}$, $\{\varepsilon_t\}_{t\in T^{(m)}}$ are mutually independent conditional on the selected contexts, we obtain that with probability at least $1-T^{-2}-\exp(\log d - \log K\cdot|T^{(m)}|/2)$,

$$\begin{aligned}
&\frac{1}{|T^{(m)}|}\sum_{t\in T^{(m)}} (x_{t,a_t}^\top\hat{\theta}_m - x_{t,a_t}^\top\theta^\star)\varepsilon_t\\
&\leq \sum_{l=1}^d M_{n,l}\sqrt{\frac{2(\log d + 2\log T)}{|T^{(m)}|}}|\hat{\theta}_{m,l} - \theta_l^\star| \leq \frac{\lambda_m}{2}\|\hat{\theta}_m - \theta^\star\|_1.
\end{aligned}$$

With the previous two inequalities together, we obtain that

$$\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}}(x_{t,a_t}^\top\theta^\star - x_{t,a_t}^\top\hat{\theta}_m)^2 + \frac{\lambda_m}{2}\|\hat{\theta}_m-\theta^\star\|_1 \le \lambda_m(\|\theta^\star\|_1 - \|\hat{\theta}_m\|_1 + \|\hat{\theta}_m-\theta^\star\|_1). \tag{D.7}$$

Define $S_0 = \mathrm{supp}(\theta^\star)$. An immediate result of (D.7) is that

$$\frac{1}{2}\|\hat{\theta}_m-\theta^\star\|_1 \le \|\theta^\star(S_0)\|_1 - \|\hat{\theta}_m(S_0)\|_1 + \|\hat{\theta}_m(S_0)-\theta^\star(S_0)\|_1$$
$$\Rightarrow \|\hat{\theta}_m(S_0^c)-\theta^\star(S_0^c)\|_1 \le 3\|\hat{\theta}_m(S_0)-\theta^\star(S_0)\|_1.$$

Before proving the final result, we state the following lemma from Bickel et al. (2009) that links the restricted eigenvalues to the condition for recovering sparse signals, where we slightly modify the notation in our presentation.

Lemma D.1 (Bickel et al. 2009). *Fix a matrix A. Assume that there exists an integer r, such that $r \ge s_0$ and $s_0 + r \le d$, such that*

$$\kappa \triangleq \sqrt{\phi_{\min}(s_0+r,A)}\left(1-3\sqrt{\frac{s_0\phi_{\max}(r,A)}{r\phi_{\min}(s_0+r,A)}}\right) > 0.$$

Then

$$\min\left\{\frac{v^\top A v}{\|v(S)\|_2^2} : S\subset[d], |S|\le s_0, v\ne 0, \|v(S^c)\|_1 \le 3\|v(S)\|_1\right\} > 0, \tag{D.8}$$

$$\min\left\{\frac{v^\top A v}{\|v(\tilde{S})\|_2^2} : S\subset[d], |S|\le s_0, v\ne 0, \|v(S^c)\|_1 \le 3\|v(S)\|_1\right\} = \kappa^2 > 0, \tag{D.9}$$

where $\tilde{S}$ is the union of S and the set of r largest in absolute value coordinates of v outside S.

Now take $r = \frac{1152 s_0 \log K}{\gamma(K)\rho(K)}$. By construction, $r \ge s_0$ and Assumption 4 ensures $s_0 + r \le d$. Using Lemma 5 with $s = s_0 + r$, we have with probability at least $1 - 2M^2\cdot\exp(O(s_0\frac{\log K\log d}{\gamma(K)\rho(K)} - \Omega(\rho^2(K)\cdot\sqrt{Ts_0}/M))$,

$$\phi_{\max}\left(s_0+r, \frac{A_m}{|T^{(m)}|}\right) \le 16\log(K), \quad \phi_{\min}\left(s_0+r, \frac{A_m}{|T^{(m)}|}\right) \ge \frac{\gamma(K)\rho(K)}{4}.$$

On the previous event, we have

$$\frac{9s_0\phi_{\max}(r, A_m/|T^{(m)}|)}{r\phi_{\min}(s_0+r, A_m/|T^{(m)}|)} \le \frac{576 s_0\log K}{r\gamma(K)\rho(K)} = \frac{1}{2},$$

and

$$\kappa = \sqrt{\phi_{\min}\left(s_0+r,\frac{A_m}{|T^{(m)}|}\right)}\left(1-3\sqrt{\frac{s_0\phi_{\max}(r,A_m/|T^{(m)}|)}{r\phi_{\min}(s_0+r,A_m/|T^{(m)}|)}}\right) \ge \frac{\sqrt{\gamma(K)\rho(K)}}{2}\cdot\left(1-\frac{\sqrt{2}}{2}\right) > 0.$$

By Lemma D.1, both (D.8) and (D.9) hold, and consequently

$$\frac{1}{|T^{(m)}|}\sum_{t\in T^{(m)}}(x_{t,a_t}^\top\theta^\star - x_{t,a_t}^\top\hat{\theta}_m)^2 \ge \kappa^2\|\theta^\star(\tilde{S}_0) - \hat{\theta}_m(\tilde{S}_0)\|_2^2. \tag{D.10}$$

Additionally by (D.7),

$$\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}}(x_{t,a_t}^\top\theta^\star - x_{t,a_t}^\top\hat{\theta}_m)^2 \le 2\lambda_m\|\hat{\theta}_m(S_0)-\theta^\star(S_0)\|_1 \overset{(a)}{\le} 2\lambda_m\sqrt{s_0}\|\hat{\theta}_m(S_0)-\theta^\star(S_0)\|_2 \le 2\lambda_m\sqrt{s_0}\|\hat{\theta}_m(\tilde{S}_0)-\theta^\star(\tilde{S}_0)\|_2, \tag{D.11}$$

where step (a) is due to the Cauchy-Schwarz inequality. Combining (D.10) and (D.11) yields

$$\|\theta^\star(\tilde{S}_0) - \hat{\theta}_m(\tilde{S}_0)\|_2 \le \frac{4\lambda_m\sqrt{s_0}}{\kappa^2}.$$

Observe that the kth largest coordinates of $|\theta^\star(S_0^c) - \hat{\theta}_m(S_0^c)|$ is bounded by $\|\theta^\star(S_0^c) - \hat{\theta}_m(S_0^c)\|_1/k$, and consequently,

$$\|\theta^\star(\tilde{S}_0^c) - \hat{\theta}_m(\tilde{S}_0^c)\|_2^2 \le \|\theta^\star(S_0^c) - \hat{\theta}_m(S_0^c)\|_1^2\sum_{k=r+1}^{d}\frac{1}{k^2} \le \frac{1}{r}\|\theta^\star(S_0^c) - \hat{\theta}_m(S_0^c)\|_1^2 \le \frac{9}{r}\|\theta^\star(S_0) - \hat{\theta}_m(S_0)\|_1^2 \le \frac{9s_0}{r}\|\theta^\star(\tilde{S}_0) - \hat{\theta}_m(\tilde{S}_0)\|_2^2,$$

where the last inequality follows from the Cauchy-Schwarz inequality. A result of the previously inequality is that

$$\|\theta^\star - \hat{\theta}_m\|_2 \overset{(a)}{\le} \left(1+3\sqrt{\frac{s_0}{r}}\right)\|\theta^\star(\tilde{S}_0) - \hat{\theta}_m(\tilde{S}_0)\|_2 \le \left(1+3\sqrt{\frac{s_0}{r}}\right)\frac{4\lambda_m\sqrt{s_0}}{\kappa^2}.$$

Finally taking a union bound, we conclude that with probability at least $1 - MT^{-2} - M\exp(\log d - \log K\cdot\Omega(\sqrt{Ts_0}/M)) - 2M^2\exp(O(s_0\frac{\log K\log d}{\gamma(K)\rho(K)} - \Omega(\rho^2(K)\cdot\sqrt{Ts_0}/M))$, for any $m\in[M]$,

$$\|\hat{\theta}_m - \theta^\star\|_2 \le \frac{800\sqrt{2}}{\gamma(K)\rho(K)}\cdot\sqrt{s_0 M}\cdot\sqrt{\frac{\log K(\log d + 2\log T)}{t_m}}.$$

Appendix E. Auxiliary Lemmas

Lemma E.1. *Suppose that $\theta \sim \mathrm{Unif}(\mathbb{S}^{s_0-1})$, then the moment of $|\theta_1|$ can be computed:*

$$\mathbb{E}|\theta_1|^p = \begin{cases} \frac{2\Gamma(\frac{s_0}{2}+1)}{\sqrt{\pi}s_0\Gamma(\frac{s_0+1}{2})} & p=1,\\ \frac{1}{s_0} & p=2,\\ \frac{4\Gamma(\frac{s_0}{2}+1)}{\sqrt{\pi}s_0(s_0+1)\Gamma(\frac{s_0+1}{2})} & p=3,\\ \frac{3}{s_0(s_0+2)} & p=4. \end{cases}$$

Moreover, we have that $\frac{2}{5\sqrt{s_0}} \le \mathbb{E}|\theta_1| \le \frac{2}{\sqrt{s_0}}$.

Proof of Lemma E.1. The density of θ is $f(\theta) = f(\theta_2,\dots,\theta_{s_0}) = \left(\frac{s_0\pi^{s_0/2}}{\Gamma(\frac{s_0}{2}+1)}\right)^{-1}\frac{2}{\sqrt{1-\theta_2^2-\cdots-\theta_{s_0}^2}}\cdot\mathbf{1}(\sum_{l=2}^{s_0}\theta_l^2\le 1)$, where $\Gamma(x) = \int_0^\infty s^{x-1}e^{-s}ds$ is the Gamma function. To compute the integrals, we leverage the spherical coordinates

$$\begin{cases} \theta_2 = r\cos\phi_1,\\ \theta_3 = r\sin\phi_1\cos\phi_2,\\ \vdots\\ \theta_{s_0-1} = r\sin\phi_1\sin\phi_2\cdots\sin\phi_{s_0-3}\cos\phi_{s_0-2},\\ \theta_{s_0} = r\sin\phi_1\sin\phi_2\cdots\sin\phi_{s_0-3}\sin\phi_{s_0-2}. \end{cases}$$

Then by direct calculation,

$$\begin{aligned}\mathbb{E}|\theta_1| &= \left(\frac{s_0\pi^{s_0/2}}{\Gamma(\frac{s_0}{2}+1)}\right)^{-1}\int_{\sum_{l=2}^{s_0}\theta_l^2\le 1} 2d\theta_2\ldots d\theta_{s_0}\\ &= 2\left(\frac{s_0\pi^{s_0/2}}{\Gamma(\frac{s_0}{2}+1)}\right)^{-1}\int_0^1\int_0^\pi\cdots\int_0^{2\pi} r^{s_0-2}\sin^{s_0-3}\phi_1\sin^{s_0-4}\phi_2\\ &\quad\cdots\sin\phi_{s_0-3}drd\phi_1d\phi_2\cdots d\phi_{s_0-2}\\ &= 2\left(\frac{s_0\pi^{s_0/2}}{\Gamma(\frac{s_0}{2}+1)}\right)^{-1}\cdot\frac{1}{s_0-1}\cdot\frac{\Gamma(\frac{s_0-2}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{s_0-1}{2})}\cdot\frac{\Gamma(\frac{s_0-3}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{s_0-2}{2})}\\ &\quad\cdots\frac{\Gamma(1)\Gamma(\frac{1}{2})}{\Gamma(\frac{3}{2})}\cdot 2\pi\\ &= \frac{2}{\sqrt{\pi}s_0}\frac{\Gamma(\frac{s_0+2}{2})}{\Gamma(\frac{s_0+1}{2})} = \begin{cases}\dfrac{1}{2^{s_0-1}}\begin{pmatrix}s_0-1\\(s_0-1)/2\end{pmatrix}, & \text{if } s_0 \text{ is odd},\\ \dfrac{2^{s_0+1}}{\pi s_0}\begin{pmatrix}s_0\\s_0/2\end{pmatrix}^{-1}, & \text{if } s_0 \text{ is even}.\end{cases}\end{aligned}$$

From the Sterling's formula, we arrive at $\frac{2}{5\sqrt{s_0}}\le\mathbb{E}|\theta_1|\le\frac{2}{\sqrt{s_0}}$. Similarly, we can compute the higher moments of $|\theta_1|$. As for the second moment,

$$\mathbb{E}|\theta_1|^2 = 4\left(\frac{s_0\pi^{s_0/2}}{\Gamma(\frac{s_0}{2}+1)}\right)^{-1}\cdot\frac{\Gamma(\frac{s_0-1}{2})\Gamma(\frac{3}{2})}{2\Gamma(\frac{s_0}{2}+1)}\cdot\frac{\pi^{s_0-1/2}}{\Gamma(\frac{s_0-1}{2})} = \frac{1}{s_0}.$$

For the third moment,

$$\begin{aligned}\mathbb{E}|\theta_1|^3 &= 4\left(\frac{s_0\pi^{s_0/2}}{\Gamma(\frac{s_0}{2}+1)}\right)^{-1}\cdot\frac{2}{s_0^2-1}\cdot\frac{\pi^{s_0-1/2}}{\Gamma(\frac{s_0-1}{2})}\\ &= \frac{4}{\sqrt{\pi}s_0(s_0+1)}\frac{\Gamma(\frac{s_0+2}{2})}{\Gamma(\frac{s_0+1}{2})}.\end{aligned}$$

For the fourth moment,

$$\begin{aligned}\mathbb{E}|\theta_1|^4 &= 4\left(\frac{s_0\pi^{s_0/2}}{\Gamma(\frac{s_0}{2}+1)}\right)^{-1}\cdot\frac{3\sqrt{\pi}\Gamma(\frac{s_0-1}{2})}{4(s_0+2)\Gamma(\frac{s_0}{2}+1)}\cdot\frac{\pi^{s_0-1/2}}{\Gamma(\frac{s_0-1}{2})}\\ &= \frac{3}{s_0(s_0+2)}.\end{aligned}$$

Appendix F. Parallel Results Under Model-P

In this section, we collect parallel results under Model-P. In what follows, Section F.1 formulates the problem, stating the conditions and assumptions. Section F.2 establishes the regret lower bound of $\Omega\Big(c\cdot\max\Big\{M^{-2}2^{-M}\cdot\sqrt{Ts_0}(T/s_0)^{\frac{1}{2(2^M-1)}}, \sqrt{Ts_0}\Big\}\Big)$. Section F.3 presents a matching upper bound of the regret under the set of (generic) assumptions stated in Section F.1. In particular, we verify the assumptions in the case of two arms, whereas that of $K>2$ needs more delicate analysis—we leave that for future work.

F.1. Problem Formulation

Recall that under Model-P, we have a set of K parameters $\{\theta_1^\star,\ldots,\theta_K^\star\}$; when action $a\in[K]$ is chosen, a reward $r_{t,a} = x_t^\top\theta_a^\star+\xi_t$ is incurred, where $\{\xi_t\}_{t=0}^\infty$ is a sequence of i.i.d. zero-mean 1-sub-Gaussian random variables. We assume $\|\theta_a^\star\|_2\le 1$ for all $a\in[K]$, and the contexts x_t are i.i.d. drawn. Assumptions F.1–F.4 are parallel to Assumptions 1–4 under Model-C.

Assumption F.1 (Sub-Guassianity). *The marginal distribution of x_t is 1-sub-Gaussian.*

Assumption F.2 (Diverse Covariate). *There are (possibly K-dependent) positive constants $\gamma(K)$ and $\rho(K)$, such that for any $\{\theta_a\}_{a\in[K]}$, any unit vector $v\in\mathbb{R}^d$ and any $a\in[K]$, there is $\mathbb{P}(v^\top x_t x_t^\top v\cdot\mathbf{1}\{a^*=a\}\ge\gamma(K))\ge\rho(K)$, where $a^* = \arg\max_{a\in[K]} x_t^\top\theta_a$.*

Assumption F.3 (Sparsity in High Dimension). *The linear contextual bandits have high-dimensional contexts $d=\mathbf{Poly}(T)$ and sparse parameters: There exists some $\varepsilon>0$ such that $\|\theta_a^\star\|_0\le s_0 = O(T^{1-\varepsilon})$ for all $a\in[K]$.*

Assumption F.4 (Not Many Actions). *The number of actions K satisfies $\frac{\log K}{\gamma(K)\rho(K)} = O(d/s_0)$ and $\frac{\log K}{\gamma(K)\rho^3(K)} = O(\sqrt{T^{1-\varepsilon}/s_0})$.*

The following lemma establishes the sufficient condition for Assumption F.2 for $K=2$.

Lemma F.1. *When $K=2$, suppose both of the following conditions hold:*

1. *There exists a constant $\Lambda>0$ such that $\lambda_{\min}(\mathbb{E}[x_t x_t^\top])\ge\Lambda$; for any unit vector $v\in\mathbb{R}^d$ there is $v^\top\mathbb{E}[x_t x_t^\top]v\ge\Gamma$;*
2. *There exists a constant $\nu>0$ such that the distribution of x_t satisfies $p(x_t)\ge\nu\cdot p(-x_t)$.*

Then Assumption F.2 *holds with $\gamma(K)=\Lambda/2$ and $\rho(K) = \nu\Lambda^2/128$.*

Proof of Lemma F.1. For any unit vector $v\in\mathbb{R}^d$, and $a\in\{1,2\}$, we have

$$\mathbb{P}\left((v^\top x_t)^2\cdot\mathbf{1}\{a^*=a\}\ge\frac{\Lambda}{2}\right)\ge\nu\cdot\mathbb{P}\left((v^\top x_t)^2\cdot\mathbf{1}\{a^*=3-a\}\ge\frac{\Lambda}{2}\right).$$

As a result,

$$\begin{aligned}\mathbb{P}\left((v^\top x_t)^2\cdot\mathbf{1}\{a^*=a\}\ge\frac{\Lambda}{2}\right) &\ge \frac{\nu}{2}\cdot\mathbb{P}\left((v^\top x_t)^2\ge\frac{\Lambda}{2}\right)\\ &\ge \frac{\nu}{2}\cdot\mathbb{P}\left((v^\top x_t)^2\ge\frac{1}{2}v^\top\mathbb{E}[x_t x_t^\top]v\right)\\ &\ge \frac{\nu\Lambda^2}{128},\end{aligned}$$

where the last inequality is due to the Paley-Zygmund inequality. □

F.2. Regret Lower Bound

Theorem F.1. *Under* Model-P, *consider the setting where $K=\log(T/s_0)$ and the context $x_t\sim\mathcal{N}(0,I_d)$ for any $t\in[T]$. For any $M\le T$ and for any dynamic batch learning algorithm **Alg**, we have*

$$\begin{aligned}&\sup_{\{\theta_a^\star\}_{a\in[K]}:\|\theta_a^\star\|_2\le 1,\|\theta_a^\star\|_1\le s_0}\mathbb{E}_{\{\theta_a^\star\}_{a\in[K]}}[R_T(\mathbf{Alg})]\\ &\ge c\cdot\max\left(M^{-2}2^{-M}\cdot\sqrt{Ts_0}\cdot\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},\sqrt{Ts_0}\right),\end{aligned}\tag{F.1}$$

where $\mathbb{E}_{\{\theta_a^\star\}_{a\in[K]}}$ denotes taking expectation w.r.t. the distribution based on the set of parameters $\{\theta_a^\star\}_{a\in[K]}$, and $c>0$ is a numerical constant independent of (T,M,d,s_0).

The proof of Theorem F.1 is similar to that of Theorem 1. We define for any $m \in [M]$,

$$\Delta_m = \frac{1}{48 \cdot M^2 \cdot 2^M} \cdot \left(\frac{T}{s_0}\right)^{-\frac{1-2^{1-m}}{2(1-2^{-M})}}, \qquad T_m = \left\lfloor s_0 \cdot \left(\frac{T}{s_0}\right)^{\frac{1-2^{-m}}{1-2^{-M}}} \right\rfloor.$$

We consider $K = 2^M$ arms and construct a prior Q for $\{\theta_a^\star\}_{a\in[K]}$ in the following way: draw $\overline{\theta}_1, \ldots, \overline{\theta}_M$ independently from $\mathsf{Unif}(\mathbb{S}^{s_0-1})$. Given $a \in [K]$, we can uniquely write $a = 1 + \sum_{m=1}^M a_m \cdot 2^{m-1}$, where $a_m \in \{0,1\}$. We then let $\tilde{\theta}_a = \sum_{m=1}^M (-1)^{a_m} \cdot \Delta_m \overline{\theta}_m$ and $\theta_a^\star$ be a d-dimensional vector whose first s_0 coordinates coincide with $\tilde{\theta}_a$ and the remaining zeros. Moving on, we let $u_t = x_t(S)/\|x_t(S)\|$.

For notational simplicity, we let $\Theta = (\theta_1, \ldots, \theta_K)$ and correspondingly $\Theta^\star = (\theta_1^\star, \ldots, \theta_K^\star)$. Then

$$\sup_{\Theta^\star:\|\theta_a^\star\|_2\le 1, \|\theta_a^\star\|_0 \le s_0, a\in[K]} \mathbb{E}_{\Theta^\star}[R_T(\mathbf{Alg})] \ge \mathbb{E}_Q \mathbb{E}_\Theta[R_T(\mathbf{Alg})] = \sum_{t=1}^T \mathbb{E}_Q \mathbb{E}_x \mathbb{E}_{P^t_{\Theta,x}} \left[\max_{a\in[K]} x_t^\top \theta_a - x_t^\top \theta_{a_t}\right].$$

Given any $m \in [M]$ and any $t \in \{T_{m-1}+1, \ldots, T_m\}$, define $\mathcal{A}_m = \{a \in [K] : a_m = 0\}$ and

$$\begin{aligned}
&\mathbb{E}_Q \mathbb{E}_x \mathbb{E}_{P^t_{\Theta,x}} \left[\max_{a\in[K]} x_t^\top \theta_a - x_t^\top \theta_{a_t}\right] \\
&= \mathbb{E}_Q \mathbb{E}_x \mathbb{E}_{P^t_{\Theta,x}} \left[\sum_{a\in[K]} \mathbf{1}\{a_t = a\} \cdot \left(\max_{a'\in[K]} x_t^\top \theta_{a'} - x_t^\top \theta_a\right)\right] \\
&= \mathbb{E}_Q \mathbb{E}_x \mathbb{E}_{P^t_{\Theta,x}} \left[\sum_{a\in\mathcal{A}_m} \mathbf{1}\{a_t = a\} \cdot \left(\max_{a'\in[K]} x_t^\top \theta_{a'} - x_t^\top \theta_a\right)\right. \\
&\quad \left. + \mathbf{1}\{a_t = a + 2^{m-1}\} \cdot \left(\max_{a'\in[K]} x_t^\top \theta_{a'} - x_t^\top \theta_{a+2^{m-1}}\right)\right] \\
&\ge \mathbb{E}_Q \mathbb{E}_x \mathbb{E}_{P^t_{\Theta,x}} \left[\sum_{a\in\mathcal{A}_m} \mathbf{1}\{a_t = a\} \cdot \left(\max_{a'\in\{a, a+2^{m-1}\}} x_t^\top \theta_{a'} - x_t^\top \theta_a\right)\right. \\
&\quad \left. + \mathbf{1}\{a_t = a + 2^{m-1}\} \cdot \left(\max_{a'\in\{a, a+2^{m-1}\}} x_t^\top \theta_{a'} - x_t^\top \theta_{a+2^{m-1}}\right)\right] \\
&\ge 2\Delta_m \cdot \mathbb{E}_Q \mathbb{E}_x \mathbb{E}_{P^t_{\Theta,x}} \left[\sum_{a\in\mathcal{A}_m} \mathbf{1}\{a_t = a\} \cdot \left(x_t(S)^\top \overline{\theta}_m\right)_- \right. \\
&\quad \left. + \mathbf{1}\{a_t = a + 2^{m-1}\} \cdot \left(x_t(S)^\top \overline{\theta}_m\right)_+\right] \\
&= 2\Delta_m \cdot \mathbb{E}_Q \mathbb{E}_x \mathbb{E}_{P^t_{\Theta,x}} \left[\mathbf{1}\{a_t \in \mathcal{A}_m\} \cdot \left(x_t(S)^\top \overline{\theta}_m\right)_- \right. \\
&\quad \left. + \mathbf{1}\{a_t \in \mathcal{A}_m^c\} \cdot \left(x_t(S)^\top \overline{\theta}_m\right)_+\right]. \qquad \text{(F.2)}
\end{aligned}$$

We define two new measures Θ via

$$\frac{dQ^+_{m,t}}{dQ}(\Theta) = \frac{(x_t(S)^\top \overline{\theta}_m)_+}{Z_m(x_t)}, \qquad \frac{dQ^-_{m,t}}{dQ}(\Theta) = \frac{(x_t(S)^\top \overline{\theta}_m)_-}{Z_m(x_t)},$$

where $Z_m(x_t) = \mathbb{E}_Q[(x_t(S)^\top \overline{\theta}_m)_+] = \mathbb{E}_Q[(x_t(S)^\top \overline{\theta}_m)_-]$ is the common normalizing constant. With the new notation, we can write

$$\text{(F.2)} = 2\Delta_m \cdot \mathbb{E}_x \left[Z_m(x_t) \cdot \left(\mathbb{E}_{P^t_{\Theta,x}\circ Q^-_{m,t}}[\mathbf{1}\{a_t \in \mathcal{A}_m\}] + \mathbb{E}_{P^t_{\Theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{a_t \in \mathcal{A}_m^c\}]\right)\right], \qquad \text{(F.3)}$$

where $P^t_{\Theta,x} \circ Q^+_{m,t}$ (respectively, $P^t_{\Theta,x} \circ Q^-_{m,t}$) is a mixed distribution: Θ is drawn from $Q^+_{m,t}$ (respectively, $Q^-_{m,t}$) and observed rewards are then drawn from $P^t_{\Theta,x}$.

F.2.1. Regret Lower Bound When a Bad Event Happens with Large Probability.

As before, the regret is large when a bad event B_m ($t_{m-1} \le T_{m-1} < T_m \le t_m$) is likely to happen under the prior.

Lemma F.2. *If there exists $m \in [M]$, such that*

$$\sum_{t=T_{m-1}+1}^{T_m} \mathbb{E}_x[Z_m(x_t) \cdot \mathbb{E}_{P_{\Theta,x}\circ Q^+_{m,t}}[\mathbf{1}\{A_m\}]] \ge \frac{T_m - T_{m-1}}{8 \cdot M^2 \cdot 2^M}, \qquad \text{(F.4)}$$

then there eixsts a numerical constant $c > 0$, independent of (T, M, d, s_0), such that,

$$\sup_{\theta^\star:\|\theta^\star\|_2\le 1, \|\theta^\star\|_0\le s_0} \mathbb{E}_{\theta^\star}[R_T(\mathbf{Alg})] \ge \frac{c}{M^2 \cdot 2^M} \cdot \sqrt{T s_0} \left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}}.$$

For any $m \in [M]$

$$\text{(F.3)} \ge 2\Delta_m \cdot \mathbb{E}_x[Z_m(x_t) \cdot (1 - \mathsf{TV}(P^t_{\Theta,x} \circ Q^+_{m,t}, P^t_{\Theta,x} \circ Q^-_{m,t}))], \qquad \text{(F.5)}$$

where the inequality is due to $P(A) + Q(A^c) \ge 1 - \mathsf{TV}(P, Q)$. Previously,

$$\begin{aligned}
&1 - \mathsf{TV}(P^t_{\Theta,x} \circ Q^-_{m,t}, P^t_{\Theta,x} \circ Q^+_{m,t}) \\
&\overset{(a)}{\ge} 1 - \mathsf{TV}(P^{T_m}_{\Theta,x} \circ Q^-_{m,t}, P^{T_m}_{\Theta,x} \circ Q^+_{m,t}) \\
&= \int \min(dP^{T_m}_{\Theta,x} \circ Q^-_{m,t}, dP^{T_m}_{\Theta,x} \circ Q^+_{m,t}) \\
&\ge \int_{B_m} \min(dP^{T_m}_{\Theta,x} \circ Q^-_{m,t}, dP^{T_m}_{\Theta,x} \circ Q^+_{m,t}) \\
&\overset{(b)}{=} \int_{B_m} \min(dP^{T_{m-1}}_{\Theta,x} \circ Q^-_{m,t}, dP^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t}), \qquad \text{(F.6)}
\end{aligned}$$

where step (a) is due to the data-processing inequality, and step (b) follows from the fact that on the event B_m, there is $P^{T_m}_{\Theta,x} = P^{T_{m-1}}_{\Theta,x}$. Next,

$$\begin{aligned}
\text{(F.6)} &= \frac{1}{2} \int_{B_m} dP^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t} + dP^{T_{m-1}}_{\Theta,x} \circ Q^-_{m,t} - |dP^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t} - dP^{T_{m-1}}_{\Theta,x} \circ Q^-_{m,t}| \\
&= \frac{1}{2}\left(P^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t}(B_m) + P^{T_{m-1}}_{\Theta,x} \circ Q^-_{m,t}(B_m)\right) - \mathsf{TV}(dP^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t}, dP^{T_{m-1}}_{\Theta,x} \circ Q^-_{m,t}) \\
&\ge P^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t}(B_m) - \frac{3}{2}\mathsf{TV}(dP^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t}, dP^{T_{m-1}}_{\Theta,x} \circ Q^-_{m,t}).
\end{aligned}$$

By Pinsker's inequality,

$$\mathsf{TV}(dP^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t}, dP^{T_{m-1}}_{\Theta,x} \circ Q^-_{m,t}) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(dP^{T_{m-1}}_{\Theta,x} \circ Q^+_{m,t} \| dP^{T_{m-1}}_{\Theta,x} \circ Q^-_{m,t})}.$$

Recall that $u_t = x_t(S)/\|x_t(S)\|_2$ and under Q,

$$(\overline{\theta}_1,\dots,\overline{\theta}_m,\dots,\overline{\theta}_M) \overset{\mathrm{d}}{=} \left(\overline{\theta}_1,\dots,\overline{\theta}_m - 2(u_t^\top \overline{\theta}_m)u_t,\dots,\overline{\theta}_M\right).$$

Let $\tilde{\Theta}^{(m)} = \{\tilde{\theta}_a^{(m)}\}_{a\in[K]}$ denote the set of 2^M arms induced by $(\overline{\theta}_1,\dots,\overline{\theta}_m - 2(u_t^\top \overline{\theta}_m)u_t,\dots,\overline{\theta}_M)$:

$$\tilde{\theta}_a^{(m)} = (-1)^{a_1}\Delta_1\cdot\overline{\theta}_1 + \cdots + (-1)^{a_m}\Delta_m\cdot(\overline{\theta}_m - 2(u_t^\top\overline{\theta}_m)\cdot u_t) + \cdots + (-1)^{a_M}\Delta_M\cdot\overline{\theta}_M.$$

Then $\Theta \sim Q_{m,t}^+$ if and only if $\tilde{\Theta}^{(m)} \sim Q_{m,t}^-$. Consequently,

$$\begin{aligned} &D_{\mathrm{KL}}(P_{\Theta,x}^{T_{m-1}}\circ Q_{m,t}^+ \,\|\, P_{\Theta,x}^{T_{m-1}}\circ Q_{m,t}^-) \\ &= D_{\mathrm{KL}}\left(P_{\Theta,x}^{T_{m-1}}\circ Q_{m,t}^+ \,\|\, P_{\tilde{\Theta}^{(m)},x}^{T_{m-1}}\circ Q_{m,t}^+\right) \\ &\le \mathbb{E}_{Q_{m,t}^+}\left[D_{\mathrm{KL}}\left(P_{\Theta,x}^{T_{m-1}}\|P_{\tilde{\Theta}^{(m)},x}^{T_{m-1}}\right)\right], \end{aligned} \tag{F.7}$$

where the last inequality is due to the joint convexity of the KL divergence (see Lemma A.4). By direct computation,

$$\begin{aligned} \text{(F.7)} &= \frac{1}{2}\mathbb{E}_{Q_{m,t}^+}\left[\sum_{\tau=1}^{T_{m-1}}\left(x_\tau^\top\theta_{a_\tau} - x_\tau^\top\tilde{\theta}_{a_\tau}^{(m)}\right)^2\right] \\ &= 2\Delta_m^2\sum_{\tau=1}^{T_{m-1}}\left(x_\tau(S)^\top u_t\right)^2\cdot\mathbb{E}_{Q_{m,t}^+}[(u_t^\top\overline{\theta}_m)^2] \\ &\le 4\frac{\Delta_m^2}{s_0^{3/2}}\frac{\|x_t(S)\|_2}{Z_m(x_t)}\cdot\sum_{\tau=1}^{T_{m-1}}\left(x_\tau(S)^\top u_t\right)^2. \end{aligned}$$

The last inequality is because

$$\begin{aligned} \mathbb{E}_{Q_{m,t}^+}[(u_t^\top\overline{\theta}_m)^2] &= \frac{\mathbb{E}_Q[(x_t(S)^\top\overline{\theta}_m)_+\cdot(u_t^\top\overline{\theta}_m)^2]}{Z_m(x_t)} \\ &= \frac{1}{2}\cdot\frac{\|x_t(S)\|_2}{Z_m(x_t)}\cdot\mathbb{E}_Q[|u_t^\top\overline{\theta}_m|^3] \\ &= \frac{1}{2}\cdot\frac{\|x_t(S)\|_2}{Z_m(x_t)}\cdot\mathbb{E}_Q[|\overline{\theta}_{m,1}|^3] \le 2\cdot\frac{\|x_t(S)\|_2}{Z_m(x_t)}\cdot s_0^{-3/2}. \end{aligned}$$

Using the previous expressions,

$$\begin{aligned} \text{(F.3)} &\ge 2\Delta_m\cdot\left(\mathbb{E}_x[Z_m(x_t)\cdot P_{\Theta,x}^{T_{m-1}}\circ Q_{m,t}^+(A_m)]\right. \\ &\quad \left. -\frac{3}{2}\mathbb{E}_x\left[Z_m(x_t)\sqrt{2\frac{\Delta_m^2}{s_0^{3/2}}\frac{\|x_t(S)\|_2}{Z_m(x_t)}\sum_{\tau=1}^{T_{m-1}}\left(x_\tau(S)^\top u_t\right)^2}\right]\right) \\ &\overset{(a)}{\ge} 2\Delta_m\cdot\left(\mathbb{E}_x[Z_m(x_t)\cdot P_{\Theta,x}^{T_{m-1}}\circ Q_{m,t}^+(A_m)]\right. \\ &\quad \left. -\frac{3}{2}\sqrt{2\mathbb{E}_x\left[Z_m(x_t)^2\frac{\Delta_m^2}{s_0^{3/2}}\frac{\|x_t(S)\|}{Z_m(x_t)}\sum_{\tau=1}^{T_{m-1}}\left(x_\tau(S)^\top u_t\right)^2\right]}\right) \\ &\ge 2\Delta_m\cdot\left(\mathbb{E}_x[Z_m(x_t)\cdot P_{\Theta,x}^{T_{m-1}}\circ Q_{m,t}^+(A_m)] - 3\sqrt{\frac{\Delta_m^2 T_{m-1}}{s_0}}\right) \\ &\ge 2\Delta_m\cdot\left(\mathbb{E}_x[Z_m(x_t)\cdot P_{\Theta,x}^{T_{m-1}}\circ Q_{m,t}^+(A_m)] - \frac{1}{16\cdot M^2\cdot 2^M}\right), \end{aligned}$$

where step (a) is due to Jensen's inequality and the concavity of $x\mapsto\sqrt{x}$. Thus far, we established for any $m\in[M]$ that

$$\begin{aligned} &\max_{\Theta^*}\sum_{t=1}^{T}\mathbb{E}_{\Theta^*}\left[\max_{a\in[K]}x_t^\top\theta_a - x_t^\top\theta_{a_t}\right] \\ &\ge 2\Delta_m\cdot\sum_{t=T_{m-1}}^{T_m}\left(\mathbb{E}_x[Z_m(x_t)\cdot P_{\theta,x}^{T_{m-1}}\circ Q_{m,t}^+(A_m)] - \frac{1}{16\cdot M^2\cdot 2^M}\right). \end{aligned}$$

Taking m to be the batch satisfying Condition (F.4), we finish the proof.

F.2.2. Bad Event Happens with Large Enough Probability. It remains to show that with sufficiently high probability, (6) holds.

Lemma F.3. *There exists some $m\in[M]$, such that*

$$\sum_{t=T_{m-1}+1}^{T_m}\mathbb{E}_x[Z_m(x_t)\cdot\mathbb{E}_{P_{\Theta,x}\circ Q_{m,t}^+}[\mathbf{1}\{B_m\}]] \ge \frac{T_m - T_{m-1}}{8\cdot M^2\cdot 2^M}.$$

For any $m\in[M]$, and any $t\in\{T_{m-1}+1,\dots,T_m\}$, we have

$$\mathbb{E}_x[Z_m(x_t)\cdot P_{\theta,x}^{T_{m-1}}\circ Q_{m,t}^+(B_m)] = \mathbb{E}_x\mathbb{E}_Q[\left(x_t(S)^\top\overline{\theta}_m\right)_+\cdot P_{\Theta,x}^{T_{m-1}}(B_m)]. \tag{F.8}$$

Because $B_m = \{t_{m-1}\le T_{m-1}\le T_m\le t_m\}$ is determined by $\{x_1,a_1,r_1,\dots,x_{T_{m-1}},a_{T_{m-1}},r_{T_{m-1}}\}$, $P_{\Theta,x}^{T_{m-1}}(B_m)$ is independent of $\{x_\tau\}_{\tau>T_{m-1}}$. Consequently,

$$\begin{aligned} \text{(F.8)} &= \mathbb{E}_Q\mathbb{E}_x\left[\left(x_T(S)^\top\overline{\theta}_m\right)_+\cdot P_{\Theta,x}^{T_{m-1}}(B_m)\right] \\ &= \mathbb{E}_Q\mathbb{E}_x\mathbb{E}_{P_{\Theta,x}}\left[\left(x_T(S)^\top\overline{\theta}_m\right)_+\mathbf{1}\{B_m\}\right] \\ &\ge \mathbb{E}_Q\mathbb{E}_x\mathbb{E}_{P_{\Theta,x}}\left[\min_{m'\in[M]}\left(x_T(S)^\top\overline{\theta}_{m'}\right)_+\mathbf{1}\{B_m\}\right] \\ &= \mathbb{E}_Q\mathbb{E}_x\left[\min_{m'\in[M]}(x_T(S)^\top\overline{\theta}_{m'})_+\right]\cdot\mathbb{E}_{\tilde{Q}}\mathbb{E}_{P_{\Theta,x}}[\mathbf{1}\{B_m\}\Big], \end{aligned}$$

where the new measure $\tilde{Q}$ is defined via the change of measure:

$$\frac{d\tilde{Q}}{dQ\times dP_x}(\Theta,x) = \frac{\min_{m'\in[M]}\left(x_T(S)^\top\overline{\theta}_{m'}\right)_+}{\mathbb{E}[\min_{m'\in[M]}\left(x_T(S)^\top\overline{\theta}_{m'}\right)_+]}.$$

By the definition of u_t, there is

$$\mathbb{E}_Q\mathbb{E}_x\left[\min_{m'\in[M]}\left(x_t(S)^\top\overline{\theta}_{m'}\right)_+\right] \ge \mathbb{E}_x\left[\|x_t(S)\|_2\cdot\mathbb{E}_Q\left[\min_{m'\in[M]}u_t^\top\overline{\theta}_{m'}\right]\right].$$

We can then directly compute

$$\begin{aligned} &\mathbb{E}_Q\left[\min_{m'\in[M]}(u_T^\top\overline{\theta}_{m'})_+\right] \overset{(a)}{=} \mathbb{E}_Q\left[\min_{m'\in[M]}(\overline{\theta}_{m',1})_+\right] \\ &= \int_0^\infty\mathbb{P}\left(\min_{m'\in[M]}(\overline{\theta}_{m',1})_+ > s\right)\mathrm{d}s \\ &= \int_0^\infty\mathbb{P}((\overline{\theta}_{1,1})_+ > s)^M\mathrm{d}s = \frac{1}{2^M}\int_0^\infty\mathbb{P}(|\overline{\theta}_{1,1}|^2 > s^2)^M\mathrm{d}s \\ &\overset{(b)}{\ge} \frac{1}{2^M}\int_0^{\frac{1}{2}\mathrm{B}(\frac{1}{2},\frac{s_0-1}{2})}\left(1 - \frac{2s}{\mathrm{B}(\frac{1}{2},\frac{s_0-1}{2})}\right)^M\mathrm{d}s \\ &\ge \frac{\mathrm{B}(\frac{1}{2},\frac{s_0-1}{2})}{(M+1)2^{M+1}} \ge \frac{1}{(M+1)2^{M+1}\sqrt{s_0}}. \end{aligned}$$

Previously, $\mathsf{B}(\alpha,\beta)$ is the beta function with parameters α and β: Step (a) is because $\overline{\theta}_1,\dots,\overline{\theta}_M$ are mutually independent; and step (b) follows from the fact that $\overline{\theta}_{m,1}^2$ follows the beta distribution with parameters $1/2$ and $(s_0-1)/2$. Taking expectation over x, we then have

$$\mathbb{E}_Q\mathbb{E}_x\left[\min_{m'\in[M]}(x_T(S)^\top\overline{\theta}_{m'})_+\right]\geq\frac{1}{(M+1)2^{M+2}}.$$

Furthermore, because the union of $\{B_m\}_{m\in[M]}$ is the whole space, by a union bound, we have $\sum_{m=1}^M\mathbb{E}_{\tilde{Q}}\mathbb{E}_{P_{\Theta,x}}[\mathbf{1}\{B_m\}]\geq\mathbb{E}_{\tilde{Q}}\mathbb{E}_{P_{\Theta,x}}[\mathbf{1}\{\cup_{m=1}^M B_m\}]=1$. Hence, there must exist $\overline{m}\in[M]$ such that $\mathbb{E}_{\tilde{Q}}\mathbb{E}_{P_{\Theta,x}}(B_{\overline{m}})\geq 1/M$ and

$$\sum_{t=T_{\overline{m}-1}}^{T_{\overline{m}}}\mathbb{E}_x\left[Z_{\overline{m}}(x_t)\cdot\mathbb{E}_{P_{\Theta,x}\circ Q^+_{\overline{m},t}}[\mathbf{1}\{B_{\overline{m}}\}]\right]1\geq\frac{T_{\overline{m}}-T_{\overline{m}-1}}{M(M+1)2^{M+2}}\geq\frac{T_{\overline{m}}-T_{\overline{m}-1}}{8\cdot M^2\cdot 2^M},$$

completing the proof.

F.2.3. Lower Bound for Fully Online Learning Setting.

It suffices now to show that the regret is lower bounded by the second term in (3). This is established in the following lemma.

Lemma F.4. *When $M=T$, under the setting of two independent Guassian contexts, we have (for some numerical constant c independent of T,M,d,s_0):*

$$\sup_{\Theta^*:\|\theta_a^*\|_2\leq 1,\|\theta_a^*\|_0\leq s_0,a\in\{1,2\}}\mathbb{E}_{\theta^*}[R_T(\mathbf{Alg})]\geq c\cdot\sqrt{Ts_0}.$$

As in the batched case (and with the same notation), we construct a prior Q for Θ^*: sample $\overline{\theta}$ from $\mathsf{Unif}(\mathbb{S}^{s_0-1})$; we then construct $\theta_1\in\mathbb{R}^d$ such that $\theta_1(S)=\Delta\overline{\theta}$ and $\theta_1(S^c)=0$, where $\Delta=\frac{1}{8}\sqrt{\frac{s_0}{T}}$. Finally, we let $\theta_2=-\theta_1$. Then,

$$\begin{aligned}&\sup_{\Theta^*:\|\theta_a^*\|_2\leq 1,\|\theta_a^*\|_0\leq s_0,a\in\{1,2\}}\mathbb{E}_{\Theta^*}[R_T(\mathbf{Alg})]\geq\mathbb{E}_Q\mathbb{E}_\Theta[R_T(\mathbf{Alg})]\\&=\sum_{t=1}^T\mathbb{E}_Q\mathbb{E}_x\mathbb{E}_{P^t_{\Theta,x}}\left[\max_{a\in\{1,2\}}(x_t^\top\theta_a-x_t^\top\theta_{a_t})\right]\\&=2\Delta\sum_{t=1}^T\mathbb{E}_x\left[Z(x_t)\cdot\left(\mathbb{E}_{P_{\Theta,x}\circ Q_t^-}[\mathbf{1}(a_t=1)]+\mathbb{E}_{P_{\Theta,x}\circ Q_t^+}[\mathbf{1}(a_t=2)]\right)\right],\end{aligned}\tag{F.9}$$

where we similarly define two measures via $\frac{dQ_t^-}{dQ}(\Theta)=\frac{(x_t(S)^\top\overline{\theta})_-}{Z(x_t)}$, $\frac{dQ_t^+}{dQ}(\Theta)=\frac{(x_t(S)^\top\overline{\theta})_+}{Z(x_t)}$, with $Z(x_t)=\frac{1}{2}\mathbb{E}_Q[|x_t(S)^\top\overline{\theta}|]$ being a common normalizing constant. Note that $\overline{\theta}\overset{\mathrm{d}}{=}\overline{\theta}-2(u_t^\top\overline{\theta})u_t$. Let $\tilde{\Theta}=\{\tilde{\theta}_1,\tilde{\theta}_2\}$ be the set of vectors induced by $\overline{\theta}-2(u_t^\top\overline{\theta})u_t$. Then $\Theta\sim Q_t^-$ if and only if $\tilde{\Theta}\sim Q_t^+$. Using this representation, we have

$$\text{(F.9)}\overset{(a)}{\geq}2\Delta\sum_{t=1}^T\mathbb{E}_x\left[Z(x_t)\cdot\left(1-\mathsf{TV}(P^{t-1}_{\Theta,x}\circ Q_t^-,P^{t-1}_{\Theta,x}\circ Q_t^+)\right)\right],\tag{F.10}$$

$$\overset{(b)}{\geq}2\Delta\sum_{t=1}^T\mathbb{E}_x\left[Z(x_t)\cdot\left(1-\sqrt{\frac{1}{2}D_{\mathrm{KL}}(P^{t-1}_{\Theta,x}\circ Q_t^-\|P^{t-1}_{\tilde{\Theta},x}\circ Q_t^-)}\right)\right],$$

$$\overset{(c)}{\geq}2\Delta\sum_{t=1}^T\mathbb{E}_x\left[Z(x_t)\cdot\left(1-\sqrt{\frac{1}{2}\mathbb{E}_{Q_t^-}[D_{\mathrm{KL}}(P^{t-1}_{\Theta,x}\|P^{t-1}_{\tilde{\Theta},x})]}\right)\right],\tag{F.11}$$

where step (a) follows from $P(A)+Q(A^c)\geq 1-\mathsf{TV}(P,Q)$; step (b) is by Pinsker's inequality; and step (c) is because of the joint convexity of the KL divergence. The KL divergence is then

$$\begin{aligned}D_{\mathrm{KL}}(P^{t-1}_{\Theta,x}\|P^{t-1}_{\tilde{\Theta},x})&=\frac{\Delta^2}{2}\sum_{\tau=1}^{t-1}\left(2(u_t^\top\overline{\theta})\cdot\left(u_t^\top x_\tau(S)\right)\right)^2\\&=2\Delta^2(u_t^\top\overline{\theta})^2\cdot u_t^\top\left(\sum_{\tau=1}^{t-1}x_\tau(S)x_\tau(S)^\top\right)u_t.\end{aligned}$$

Plugging in the expression of the KL divergence, we have

$$\begin{aligned}\text{(F.11)}&=2\Delta\sum_{t=1}^T\mathbb{E}_x[Z(x_t)\\&\quad\cdot\left(1-\sqrt{\Delta^2\mathbb{E}_{Q_t^-}[(u_t^\top\overline{\theta})^2]\cdot u_t^\top\left(\sum_{\tau=1}^{t-1}x_\tau(S)^\top x_\tau(S)\right)u_t}\right)]\\&\overset{(a)}{\geq}2\Delta\sum_{t=1}^T\mathbb{E}_x\left[Z(x_t)\cdot\left(1-\sqrt{\frac{5t\Delta^2}{s_0}}\right)\right]\\&\overset{(b)}{\geq}\frac{T\Delta}{5}=\frac{\sqrt{Ts_0}}{40},\end{aligned}$$

where step (a) is by taking expectation w.r.t. $\{x_\tau\}_{\tau\leq t-1}$ and Lemma E.1, and step (b) is the choice of Δ.

F.3. Regret Upper Bound

Algorithm F.1 describes a variant of the LBGL algorithm under Model-P; Theorem F.2 establishes a corresponding regret upper bound under Assumptions F.1–F.4, and Corollary F.1 gives an upper bound for the online learning problem.

Algorithm F.1 (LBGL Under Model-P)

Input Time horizon T; context dimension d; number of batches M; sparsity bound s_0.

Initialize $b=\Theta(\sqrt{T}\cdot(T/s_0)^{\frac{1}{2(2^M-1)}})$; $\hat{\theta}_0=\mathbf{0}\in\mathbb{R}^d$;

Static grid $\mathcal{T}=\{t_1,\dots,t_M\}$, with $t_1=b\sqrt{s_0}$ and $t_m=b\sqrt{t_{m-1}}$ for $t\in\{2,\dots,M\}$;

Partition each batch into M intervals evenly, that is, $(t_{m-1},t_m]=\cup_{j=1}^M T_m^{(j)}$, for $m\in[M]$.

for $m\leftarrow 1$ ***to*** M **do**

  for $t\leftarrow t_{m-1}$ ***to*** t_m **do**

    (a) Choose $a_t=\arg\max_{a\in[K]}x_t^\top\hat{\theta}_{m-1,a}$ (break ties with lower action index).

    (b) Incur reward r_{t,a_t}.

  end

  $T^{(m)}\leftarrow\cup_{m'=1}^m T_{m'}^{(m)}$;

  $\lambda_m\leftarrow 18\cdot\sqrt{\frac{\log T}{|T^{(m)}|}}$;

  Update $\hat{\theta}_{m,a}\leftarrow\arg\min_{\theta\in\mathbb{R}^d}\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}}(r_{t,a_t}-x_t^\top\theta)^2\cdot\mathbf{1}\{a_t=a\}+\lambda_m\|\theta\|_1$.

end

Theorem F.2. *Under* Model-P, Assumptions F.1–F.4 *and* $M=O(\mathrm{loglog}(T/s_0))$, *we have*

$$\sup_{\Theta^*:\|\theta_a^*\|_2\leq 1,\|\theta_a^*\|_0\leq s_0}\mathbb{E}_{\Theta^*}[R_T(\mathbf{Alg})]\leq\frac{C\cdot M^{3/2}\sqrt{\log\mathrm{Tlog}(TK)}}{\gamma(K)\rho(K)}\cdot\sqrt{Ts_0}\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},\tag{F.12}$$

where ***Alg*** *is LBGL and $C>0$ is a numerical constant independent of (T,d,M,K,s_0).*

Corollary F.1. *In the fully online learning setting ($M=T$) and under* Assumptions F.1–F.4*:*

$$\sup_{\Theta^*:\|\theta_a^*\|_2\le 1,\,\|\theta_a^*\|_0\le s_0}\mathbb{E}_{\theta^*}[R_T(\mathbf{Alg})] \le \frac{C\sqrt{\left(\operatorname{loglog}(T/s_0)\right)^3\cdot\log T\cdot\log(TK)}}{\gamma(K)\rho(K)}\cdot\sqrt{Ts_0}, \tag{F.13}$$

where $C>0$ is a numerical constant independent of (T,d,M,K,s_0).

F.3.1. Eigenvalue Conditions. We define for any $j,m\in[M]$ and $a\in[K]$ the empirical covariance matrix: $D_{m,j,a}=\sum_{t\in T_m^{(j)}}x_tx_t^\top\mathbf{1}\{a_t=a\}$ and $A_{m,a}=\sum_{m'=1}^{m}D_{m',m,a}$. Lemma F.5 shows that the restricted eigenvalues are bounded from both above and below with high probabilities.

Lemma F.5. *Suppose* Assumptions F.1–F.4 *hold. Given a sparsity parameter s, for any $j,m\in[M]$ and $a\in[K]$, with probability at least $1-2\exp\left(O(s\cdot\log\left(\frac{d}{\rho(K)\gamma(K)}\right)\right)-\Omega(\rho^2(K)\sqrt{Ts_0}/M))$,*

$$\phi_{\max}\left(s,\frac{D_{m,j,a}}{|T_m^{(j)}|}\right)\le\frac{25}{2},\qquad \phi_{\min}\left(s,\frac{D_{m,j,a}}{|T_m^{(j)}|}\right)\ge\frac{\rho(K)\cdot\gamma(K)}{4}.$$

Proof of Lemma F.5. Given a sparsity parameter s, let $\mathcal{N}(\varepsilon)$ denote the ε-net of $\mathbb{S}^{s-1}$.

Upper bound. Fixing an arbitrary s-sparse vector $v\in\mathbb{R}^d$, we let $Y_t=v^\top x_t$. For any $\delta,\mu>0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}Y_t^2\cdot\mathbf{1}\{a_t=a\}\ge 4+\delta\right)\\
&\overset{(a)}{\le}\exp\Big(-(4+\delta)\mu\cdot|T_m^{(j)}|\Big)\cdot\mathbb{E}\left[\exp\left(\sum_{t\in T_m^{(j)}}\mu\cdot Y_t^2\mathbf{1}\{a_t=a\}\right)\right]\\
&\le\exp\Big(-(4+\delta)\mu\cdot|T_m^{(j)}|\Big)\cdot\mathbb{E}\left[\exp\left(\sum_{t\in T_m^{(j)}}\mu\cdot Y_t^2\right)\right]\\
&\overset{(b)}{=}\exp\Big(-\delta\mu\cdot|T_m^{(j)}|\Big)\cdot\prod_{t\in T_m^{(j)}}\mathbb{E}\left[\exp\Big(\mu(Y_t^2-4)\Big)\right],
\end{aligned}\tag{F.14}$$

where the step (a) is a result of Markov's inequality, and step (b) is due to the independence between $\{x_t\}_{t\in T_m^{(j)}}$. Because x_t is 1-sub-Gaussian, $v^\top x_t$ is 1-sub-Gaussian. Hence, $Y_t^2-\mathbb{E}[Y_t^2]$ is $(4\sqrt{2},4)$-subexponential and $\mathbb{E}[Y_t^2]\le 4$. Using this result, we have

$$\begin{aligned}
(\text{F.14})&\le\exp\Big(-\delta\mu\cdot|T_m^{(j)}|\Big)\cdot\prod_{t\in T_m^{(j)}}\mathbb{E}\left[\exp\Big(\mu\cdot(Y_t^2-\mathbb{E}[Y_t^2])\Big)\right]\\
&\le\exp\left(-\min\left(\frac{\delta}{8},\frac{\delta^2}{64}\right)\cdot|T_m^{(j)}|\right).
\end{aligned}$$

Letting $\delta=8$ and taking a union bound over all the d-dimensional vectors whose support is in $\mathcal{N}(\varepsilon)$, we obtain that with probability at least $1-\exp(s\log d+s\log(1+2/\varepsilon)-|T_m^{(j)}|)$,

$$\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}(v^\top x_t)^2\cdot\mathbf{1}\{a_t=a\}<12,$$

for all v whose support is in $\mathcal{N}(\varepsilon)$. For an arbitrary s-sparse vector v, let $\operatorname{supp}(v)$ denote its support. Without loss of generality, suppose $|\operatorname{supp}(v)|=s$. By the definition of the ε-net, we can find $u_0\in\mathcal{N}(\varepsilon)$ such that $\|\operatorname{supp}(v)-u_0\|_2\le\varepsilon$. We then construct the d-dimensional vector u such that $u(\operatorname{supp}(v))=u_0$ and $u(\operatorname{supp}(v)^c)=0$. Then,

$$\begin{aligned}
&\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}(v^\top x_t)^2\cdot\mathbf{1}\{a_t=a\}-\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}(u^\top x_t)^2\cdot\mathbf{1}\{a_t=a\}\\
&=\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}v^\top x_tx_t^\top(v-u)\cdot\mathbf{1}\{a_t=a\}\\
&\quad-\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}u^\top x_tx_t^\top(u-v)\cdot\mathbf{1}\{a_t=a\}\\
&\le 2\varepsilon\cdot\phi_{\max}\left(s,\frac{D_{m,j,a}}{|T_m^{(j)}|}\right).
\end{aligned}$$

Taking the supremum over all s-sparse vectors v and rearranging the previous expression, we conclude that with probability at least $1-\exp(s\log d+s\log(1+2/\varepsilon)-|T_m^{(j)}|)$,

$$\phi_{\max}\left(s,\frac{D_{m,j,a}}{|T_m^{(j)}|}\right)\le\frac{12}{1-2\varepsilon}.$$

Lower bound. We again fix a s-sparse vector $v\in\mathbb{R}^d$ and let $Y_t=v^\top x_t$. For any $a\in[K]$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}Y_t^2\cdot\mathbf{1}\{a_t=a\}\le\frac{\rho(K)\gamma(K)}{2}\right)\\
&=\mathbb{P}\left(\frac{1}{T_m^{(j)}}\sum_{t\in T_m^{(j)}}\frac{Y_t^2}{\gamma(K)}\cdot\mathbf{1}\{a_t=a\}\le\frac{\rho(K)}{2}\right)\\
&\le\mathbb{P}\left(\frac{1}{T_m^{(j)}}\sum_{t\in T_m^{(j)}}\mathbf{1}\{Y_t^2\ge\gamma(K),a_t=a\}\le\frac{\rho(K)}{2}\right)\\
&=\mathbb{P}\left(\frac{1}{T_m^{(j)}}\sum_{t\in T_m^{(j)}}\mathbf{1}\{Y_t^2\ge\gamma(K),a_t=a\}-\mathbb{P}\Big(Y_t^2\ge\gamma(K),a_t=a\Big)\right.\\
&\qquad\left.\le\frac{\rho(K)}{2}-\mathbb{P}\Big(Y_t^2\ge\gamma(K),a_t=a\Big)\right)\\
&\le\mathbb{P}\left(\frac{1}{T_m^{(j)}}\sum_{t\in T_m^{(j)}}\mathbf{1}\{Y_t^2\ge\gamma(K),a_t=a\}\right.\\
&\qquad\left.-\mathbb{P}\Big(Y_t^2\ge\gamma(K),a_t=a\Big)\le-\frac{\rho(K)}{2}\right),
\end{aligned}\tag{F.15}$$

where the last inequality is due to Assumption 2. Applying Hoeffding's inequality, we obtain that

$$(39)\le\exp\left(-\frac{|T_m^{(j)}|\cdot\rho^2(K)}{2}\right).$$

Taking a union bound over all the d-dimensional vectors whose support is in $\mathcal{N}(\varepsilon)$, we have with probability at least $1-\exp(s\log d+s\log(1+2/\varepsilon)-|T_m^{(j)}|\cdot\rho^2(K)/2)$,

$$\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}}(v^\top x_t)^2\cdot\mathbf{1}\{a_t=a\}\ge\frac{\rho(K)\gamma(K)}{2},$$

for all v whose support is in $\mathcal{N}(\varepsilon)$. Conditional on the previous event, for an arbitrary s-sparse d-dimensional vector v, by the definition of an ε-net, we can find $u_0 \in \mathcal{N}(\varepsilon)$ such that $\|u_0 - \mathsf{supp}(v)\|_2 \le \varepsilon$. Let $u \in \mathbb{R}^d$ such that $u(\mathsf{supp}(v)) = u_0$ and $u(\mathsf{supp}(v)^c) = 0$. We then have

$$\begin{aligned}
&\frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}} (v^\top x_t)^2 \cdot \mathbf{1}\{a_t = a\} \\
&\ge \frac{1}{|T_m^{(j)}|}\sum_{t\in T_m^{(j)}} \left((u^\top x_t)^2 + 2(v-u)^\top x_t x_t^\top u\right)\cdot \mathbf{1}\{a_t = a\} \\
&\ge \frac{\rho(K)\gamma(K)}{2} - 2\varepsilon\phi_{\max}\left(s, \frac{D_{m,j,a}}{|T_m^{(j)}|}\right).
\end{aligned}$$

Finally, taking $\varepsilon = \min(\frac{1}{50}, \frac{\rho(K)\gamma(K)}{100})$, we have with probability at least $1 - 2\exp\left(O(s\log\frac{d}{\rho(K)\gamma(K)}) - \Omega(\rho^2(K)\sqrt{Ts_0}/M)\right)$,

$$\phi_{\max}\left(s, \frac{D_{m,j,a}}{|T_m^{(j)}|}\right) \le \frac{25}{2}, \qquad \phi_{\min}\left(s, \frac{D_{m,j,a}}{|T_m^{(j)}|}\right) \ge \frac{\rho(K)\cdot\gamma(K)}{4}.$$

F.3.2. Lasso Estimation Error With well-behaved restricted eigenvalues, Lemma F.6 leverages standard Lasso results to prove an estimation error bound for $\|\hat\theta_{m,a} - \theta_a^\star\|_2$.

Lemma F.6. *Suppose Assumptions F.1–F.4 hold. Given any $a \in [K]$ and $m \ge 2$, with probability at least $1 - 2M\exp(O(\frac{s_0}{\rho(K)\gamma(K)})\cdot \log(\frac{d}{\rho(K)\gamma(K)}) - \Omega(\sqrt{Ts_0}/M)) - 2T^{-2} - 2\exp(\log d - \Omega(\sqrt{Ts_0}/M))$,*

$$\|\theta_a - \hat\theta_{m,a}\|_2 \le \frac{2{,}048}{\rho(K)\gamma(K)}\cdot\sqrt{\frac{Ms_0\log T}{t_m}}.$$

By the definition of $\hat\theta_{m,a}$,

$$\begin{aligned}
&\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}} (r_{t,a_t} - x_t^\top\hat\theta_{m,a})^2\cdot\mathbf{1}\{a_t = a\} + \lambda_m\|\hat\theta_{m,a}\|_1 \\
&\le \frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}} (r_{t,a_t} - x_t^\top\theta_a)^2\cdot\mathbf{1}\{a_t = a\} + \lambda_m\|\theta_a\|_1.
\end{aligned}$$

Rearranging yields

$$\begin{aligned}
&\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}} (x_t^\top\theta_a - x_t^\top\hat\theta_{m,a})^2\cdot\mathbf{1}\{a_t = a\} + \lambda_m\|\hat\theta_{m,a}\|_1 \\
&\le \frac{1}{|T^{(m)}|}\sum_{t\in T^{(m)}} (x_t^\top\hat\theta_{m,a} - x_t^\top\theta_a)\cdot\varepsilon_t\cdot\mathbf{1}\{a_t = a\} + \lambda_m\|\theta_a\|_1.
\end{aligned}$$

Because of the construction of $T^{(m)}$, conditional on $\{x_t, a_t\}_{t\in T^{(m)}}$, $\{\varepsilon_t\}_{t\in T^{(m)}}$ are mutually independent:

$$\begin{aligned}
&\frac{1}{|T^{(m)}|}\sum_{t\in T^{(m)}} x_t^\top(\hat\theta_{m,a} - \theta_a)\cdot\varepsilon_t\cdot\mathbf{1}\{a_t = a\} \\
&= \frac{1}{|T^{(m)}|}\sum_{j\in[d]}(\hat\theta_{m,a,j} - \theta_{a,j})\sum_{t\in T^{(m)}} x_{t,j}\varepsilon_t\mathbf{1}\{a_t = a\} \\
&\le \frac{1}{|T^{(m)}|}\sum_{j\in[d]}|\hat\theta_{m,a,j} - \theta_{a,j}|\cdot\left|\sum_{t\in T^{(m)}} x_{t,j}\varepsilon_t\mathbf{1}\{a_t = a\}\right|. \qquad \text{(F.16)}
\end{aligned}$$

For a given $j \in [d]$ and $\delta, \delta_1 > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{|T^{(m)}|}\left|\sum_{t\in T^{(m)}} x_{t,j}\mathbf{1}\{a_t = a\}\varepsilon_t\right| \ge \delta\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\frac{1}{|T^{(m)}|}\left|\sum_{t\in T^{(m)}} x_{t,j}\mathbf{1}\{a_t = a\}\varepsilon_t\right| \ge \delta \mid \{x_{t,j}\}_{t\in T^{(m)}}\right)\right] \\
&\overset{(a)}{\le} 2\mathbb{E}\left[\exp\left(-\frac{|T^{(m)}|^2\delta^2}{2\sum_{t\in T^{(m)}} x_{t,j}^2\mathbf{1}\{a_t = a\}}\right)\right] \\
&\overset{(b)}{\le} 2\exp\left(-\frac{|T^{(m)}|\delta^2}{2(4+\delta_1)}\right) + 2\mathbb{P}\left(\frac{1}{|T^{(m)}|}\sum_{t\in T^{(m)}} x_{t,j}^2 \ge 4 + \delta_1\right),
\end{aligned}$$

where step (a) uses Hoeffding's inequality, and step (b) applies a union bound. By the assumption, x_t is 1-sub-Gaussian, and hence $x_{t,j}$ is also 1-sub-Gaussian; $x_{t,j}^2 - \mathbb{E}[x_{t,j}^2]$ is $(4\sqrt{2}, 4)$-subexponential and $\mathbb{E}[x_{t,j}^2] \le 4$. Consequently,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{|T^{(m)}|}\sum_{t\in T^{(m)}} x_{t,j}^2 \ge 4 + \delta_1\right) &\le \mathbb{P}\left(\frac{1}{|T^{(m)}|}\sum_{t\in T^{(m)}} x_{t,j}^2 - \mathbb{E}[x_{t,j}^2] \ge \delta_1\right) \\
&\le \exp\left(-\min\left(\frac{\delta_1^2}{64}, \frac{\delta_1}{8}\right)\cdot|T^{(m)}|\right).
\end{aligned}$$

Letting $\delta = 9\sqrt{\log T/|T^{(m)}|}$ and $\delta_1 = 8$ and taking a union bound over $j \in [d]$, we have with probability at least $1 - 2T^{-2} - 2\exp(\log d - |T^{(m)}|)$,

$$\frac{1}{|T^{(m)}|}\cdot\left|\sum_{t\in T^{(m)}} x_{t,j}\varepsilon_t\mathbf{1}\{a_t = a\}\right| \le 9\sqrt{\frac{\log T}{|T^{(m)}|}} = \frac{\lambda_m}{2}, \qquad \text{(F.17)}$$

for all $j \in [d]$.

On Event (F.17), (40) $\le \frac{\lambda_m}{2}\|\hat\theta_{m,a} - \theta_a\|_1$. Consequently,

$$\begin{aligned}
&\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}} (x_t^\top\theta_a - x_t^\top\hat\theta_{m,a})^2\cdot\mathbf{1}\{a_t = a\} + \frac{\lambda_m}{2}\|\hat\theta_{m,a} - \theta_a\|_1 \\
&\le \lambda_m\cdot(\|\hat\theta_{m,a} - \theta_a\|_1 + \|\theta_a\|_1 - \|\hat\theta_{m,a}\|_1).
\end{aligned}$$

Denote $S_a = \mathsf{supp}(\theta_a)$. The previous inequality yields

$$\begin{aligned}
&\frac{1}{2}\|\hat\theta_{m,a} - \theta_a\|_1 \le \|\hat\theta_{m,a}(S_a) - \theta_a(S_a)\|_1 + \|\theta_a(S_a)\|_1 - \|\hat\theta_{m,a}(S_a)\|_1 \\
&\Rightarrow \|\hat\theta_{m,a}(S_a^c) - \theta_a(S_a^c)\|_1 \le 3\|\hat\theta_{m,a}(S_a) - \theta_a(S_a)\|_1.
\end{aligned}$$

Define $B_{m,a} = \sum_{t\setminus\text{in}T^{(m)}} x_t x_t^\top\cdot\mathbf{1}\{a_t = t\}$. For any unit vector $v \in \mathbb{R}^d$,

$$v^\top\frac{B_{m,a}}{T^{(m)}}v = \sum_{j\le m}\frac{|T_m^{(j)}|}{|T^{(m)}|}\cdot v^\top\frac{D_{j,m,a}}{|T_m^{(j)}|}v.$$

Combining the previous expressions and Lemma F.5, we have with probability at least $1 - 2M\cdot\exp\left(O(s\log\frac{d}{\rho(K)\gamma(K)}\right.$ $\left.-\Omega(\rho^2(K)\cdot\sqrt{Ts_0}/M)\right)$,

$$\phi_{\max}\left(s, \frac{B_{m,a}}{|T^{(m)}|}\right) \le \frac{25}{2}, \qquad \phi_{\min}\left(s, \frac{B_{m,a}}{|T^{(m)}|}\right) \ge \frac{\rho(K)\cdot\gamma(K)}{4}.$$

We now let $r = \frac{1{,}800 s_0}{\rho(K)\gamma(K)}$ and $s = s_0 + r$. With probability at least $1 - 2M\exp(O(\frac{s_0}{\rho(K)\gamma(K)}\log\frac{d}{\rho(K)\gamma(K)}) - \Omega(\rho^2(K)\cdot\sqrt{Ts_0}/M))$,

$$\frac{9s_0\phi_{\max}\left(r, \frac{B_{m,a}}{|T^{(m)}|}\right)}{r\phi_{\min}\left(s_0 + r, \frac{B_{m,a}}{|T^{(m)}|}\right)} \le \frac{1}{4},$$

and hence

$$\kappa = \sqrt{\phi_{\min}\left(s_0 + r, \frac{B_{m,a}}{|T^{(m)}|}\right)} \cdot \left(1 - 3\sqrt{\frac{s_0\phi_{\max}\left(r, \frac{B_{m,a}}{|T^{(m)}|}\right)}{r\phi_{\max}\left(s_0 + r, \frac{B_{m,a}}{|T^{(m)}|}\right)}}\right) \ge \frac{\sqrt{\rho(K)\gamma(K)}}{4}.$$

We now make use of Lemma D.1:

$$\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}}(x_t^\top\theta_a - x_t^\top\hat\theta_{m,a})^2\cdot\mathbf{1}\{a_t = a\} \ge \frac{\rho(K)\gamma(K)}{32}\|\hat\theta_{m,a}(\tilde S_a) - \theta_a(\tilde S_a)\|_2^2. \tag{F.18}$$

Conversely,

$$\begin{aligned}&\frac{1}{2|T^{(m)}|}\sum_{t\in T^{(m)}}(x_t^\top\theta_A - x_t^\top\hat\theta_{m,a})^2\cdot\mathbf{1}\{a_t = a\}\\ &\le 2\lambda_m\cdot\|\theta_a(S_a) - \hat\theta_{m,a}(S_a)\|_1\\ &\le 2\lambda_m\sqrt{s_0}\cdot\|\theta_a(S_a) - \hat\theta_{m,a}(S_a)\|_2\\ &\le 2\lambda_m\sqrt{s_0}\cdot\|\theta_a(\tilde S_a) - \hat\theta_{m,a}(\tilde S_a)\|_2.\end{aligned} \tag{F.18}$$

Combining (F.17) and (F.18), we have

$$\|\hat\theta_{m,a}(\tilde S_a) - \theta_a(\tilde S_a)\|_2 \le \frac{64\lambda_m\sqrt{s_0}}{\rho(K)\gamma(K)}.$$

Observe that the kth largest coordinate of $|\hat\theta_{m,a}(S_a^c) - \theta_a(S_a^c)|$ is bounded by $\|\hat\theta_{m,a}(S_a^c) - \theta_a(S_a^c)\|_1/k$. Then,

$$\begin{aligned}\|\hat\theta_{m,a}(\tilde S_a^c) - \theta_a(\tilde S_a^c)\|_2^2 &\le \|\hat\theta_{m,a}(S_a^c) - \theta_a(S_a^c)\|_1^2\sum_{\ell=r+1}^{d-s_0}\frac{1}{\ell^2}\\ &\le \frac{1}{r}\cdot\|\hat\theta_{m,a}(S_a^c) - \theta_a(S_a^c)\|_1^2\\ &\le \frac{9}{r}\cdot\|\hat\theta_{m,a}(S_a) - \theta_a(S_a)\|_1^2 \le \frac{9s_0}{r}\cdot\|\hat\theta_{m,a}(S_a) - \theta_a(S_a)\|_2^2\\ &\le \frac{9s_0}{r}\cdot\|\hat\theta_{m,a}(\tilde S_a) - \theta_a(\tilde S_a)\|_2^2.\end{aligned}$$

Combining everything previously stated, we have with probability at least $1 - 2M\cdot\exp\left(O\left(\frac{s_0}{\rho(K)\gamma(K)}\log\frac{d}{\rho(K)\gamma(K)}\right) - \Omega(\rho^2(K)\cdot\sqrt{Ts_0}/M)\right) - 2T^{-2} - 2\exp(\log d - \Omega(\sqrt{Ts_0}/M))$,

$$\begin{aligned}\|\hat\theta_{m,a} - \theta_a\|_2 &\le \sqrt{1 + \frac{9s_0}{r}}\cdot\|\hat\theta_{m,a}(\tilde S_a) - \theta_a(\tilde S_a)\|_2\\ &\le \sqrt{1 + \frac{9s_0}{r}}\cdot\frac{64\lambda_m\sqrt{s_0}}{\rho(K)\gamma(K)} \le \frac{4{,}608}{\rho(K)\gamma(K)}\sqrt{\frac{Ms_0\log T}{t_m}}.\end{aligned}$$

F.3.3. Regret Analysis Given $m\in[M]$ and $t\in\{t_{m-1}+1, \ldots, t_m\}$, the instantaneous regret can be bounded as

$$\begin{aligned}&\max_{a\in[K]} x_t^\top\theta_a - x_t^\top\theta_{a_t} = \max_{a\in[K]} x_t^\top\theta_a - x_t^\top\theta_{a_t} - x_t^\top\hat\theta_{m-1,a_t} + x_t^\top\hat\theta_{m-1,a_t}\\ &\overset{(a)}{\le}\max_{a\in[K]} x_t^\top\theta_a - x_t^\top\theta_{a_t} - x_t^\top\hat\theta_{m-1,a} + x_t^\top\hat\theta_{m-1,a_t}\\ &\le 2\cdot\max_{a\in[K]}|x_t^\top(\theta_a - \hat\theta_{m-1,a})|,\end{aligned}$$

where step (a) is due to the definition of a_t. Conditional on the previous batches, $x_t^\top(\theta_a - \hat\theta_{m-1,a})$ is $\|\theta_a - \hat\theta_{m-1,a}\|_2^2$-sub-Gaussian for a given $a\in[K]$. Letting $\mathcal{H}_t = \{x_1, a_1, r_1, \ldots, x_t, a_t, r_t\}$, we have

$$\begin{aligned}&\mathbb{P}\left(2\cdot\max_{a\in[K]}|x_t^\top(\theta_a - \hat\theta_{m-1,a})| \ge 6\sqrt{\log(TK)}\cdot\max_{a\in[K]}\|\theta_a - \hat\theta_{m-1,a}\|_2\,|\,\mathcal{H}_{t_{m-1}}\right)\\ &\le\sum_{a\in[K]}\mathbb{P}\left(|x_t^\top(\theta_a - \hat\theta_{m-1,a})| \ge 3\sqrt{\log(TK)}\cdot\max_{a\in[K]}\|\theta_a - \hat\theta_{m-1,a}\|_2\,|\,\mathcal{H}_{t_{m-1}}\right)\\ &\le\sum_{a\in[K]}\mathbb{P}\left(|x_t^\top(\theta_a - \hat\theta_{m-1,a})| \ge 3\sqrt{\log(TK)}\cdot\|\theta_a - \hat\theta_{m-1,a}\|_2\,|\,\mathcal{H}_{t_{m-1}}\right) \le \frac{1}{T^4},\end{aligned}$$

where the last inequality is due to the Chernoff bound. Applying Lemma F.6 and a union bound over $t\in\{t_{m-1}+1, \ldots, t_m\}$ and $a\in[K]$, for $m\ge 2$, with probability at least $1 - T^{-3} - 2MK\exp\left(O\left(\frac{s_0}{\rho(K)\gamma(K)}\cdot\log\left(\frac{d}{\rho(K)\gamma(K)}\right)\right) - \Omega(\sqrt{Ts_0}/M)\right) - 2KT^{-2} - 2K\exp(\log d - \Omega(\sqrt{Ts_0}/M))$, we bound the regret incurred in the mth batch as

$$\begin{aligned}&\sum_{t=t_{m-1}+1}^{t_m} 2\cdot\max_{a\in[K]}|x_t^\top(\theta_a - \hat\theta_{m-1,a})| \le 12{,}288\cdot\frac{\sqrt{M\log\mathrm{T}\log(TK)}}{\rho(K)\gamma(K)}\cdot\sqrt{\frac{s_0}{t_{m-1}}}\cdot t_m\\ &\le c_1\cdot\frac{\sqrt{M\log\mathrm{T}\log(TK)}}{\rho(K)\gamma(K)}\cdot\sqrt{Ts_0}\cdot\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},\end{aligned}$$

where the last inequality follows from the choice of the grids and $c_1 > 0$ is a numerical constant. Next, for the first batch,

$$\sum_{t=1}^{t_1}\max_{a\in[K]} x_t^\top\theta_a - x_t^\top\theta_{a_t} \le 2\sum_{t=1}^{t_1}\max_{a\in[K]}|x_t^\top\theta_a|.$$

Applying a maximal sub-Gaussian inequality and taking a union bound over $t\in[t_1]$, we have with probability at least $1 - T^{-2}$,

$$\begin{aligned}\sum_{t=1}^{t_1}\max_{a\in[K]} x_t^\top\theta_a - x_t^\top\theta_{a_t} &\le 6\sqrt{\log(TK)}\cdot t_1\\ &\le c_2\sqrt{\log(TK)}\cdot\sqrt{Ts_0}\cdot\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},\end{aligned}$$

where $c_2 > 0$ is a numerical constant. Combining everything previously stated, we have with probability at least $1 - (1 + M + 2MK)\cdot T^{-2} - 2MK\exp(\log d - \Omega(\sqrt{Ts_0}/M)) - 2M^2K\exp\left(O\left(\frac{s_0}{\rho(K)\gamma(K)}\cdot\log\left(\frac{d}{\rho(K)\gamma(K)}\right)\right) - \Omega(\sqrt{Ts_0}/M)\right)$,

$$R_T(\mathbf{Alg}) \le c_3\cdot\frac{\sqrt{M^3\log\mathrm{T}\log(TK)}}{\rho(K)\gamma(K)}\cdot\sqrt{Ts_0}\cdot\left(\frac{T}{s_0}\right)^{\frac{1}{2(2^M-1)}},$$

where $c_3 > 0$ is a numerical constant independent of (T, d, M, K, s_0).

Endnotes

[1] A preprint of Bastani and Bayati (2020) occurred prior to Wang et al. (2018).

[2] In Abbasi-Yadkori (2012), a $\tilde{O}(\sqrt{s_0 dT})$ regret bound is obtained, although the contexts there can be arbitrary rather than stochastic.

[3] This result follows directly from the lower bound given in Chu et al. (2011), although our lower bound argument provides an alternative proof.

[4] This may not be the case since the various low-dimensional regime assumptions are often required to obtain the $\tilde{\Theta}(\sqrt{dT})$ regret bounds.

References

Abbasi-Yadkori Y (2012) Online learning for linearly parametrized control problems. Ph.D. Dissertation. University of Alberta, CAN. Advisor(s) Csaba Szepesvari.

Agrawal S, Goyal N (2013a) Further optimal regret bounds for Thompson sampling. Carvalho CM, Ravikumar P, eds. *Artificial Intelligence and Statistics* (PMLR, Cambridge, MA), 99–107.

Agrawal S, Goyal N (2013b) Thompson sampling for contextual bandits with linear payoffs. Dasgupta S, McAllester D, eds. *Proc. Internat. Conf. on Machine Learn.* (PMLR, Cambridge, MA), 127–135.

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learn. Res.* 3(Nov):397–422.

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Oper. Res.* 67(1):90–108.

Bastani H, Bayati M (2020) Online decision making with high-dimensional covariates. *Oper. Res.* 68(1):276–294.

Bastani H, Bastani O, Kim C (2017) Interpreting predictive models for human-in-the-loop analytics. Preprint, submitted May 23, https://arxiv.org/abs/1705.08504.

Bastani H, Bayati M, Khosravi K (2021) Mostly exploration-free algorithms for contextual bandits. *Management Sci.* 67(3): 1329–1349.

Bayati M, Braverman M, Gillam M, Mack KM, Ruiz G, Smith MS, Horvitz E (2014) Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS One* 9(10):e109264.

Belloni A, Chernozhukov V (2011) High dimensional sparse econometric models: An introduction. *Inverse Problems and High-Dimensional Estimation* (Springer, Berlin), 121–156.

Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *Rev. Econom. Stud.* 81(2):608–650.

Bertsimas D, Mersereau AJ (2007) A learning approach for interactive marketing to a customer segment. *Oper. Res.* 55(6): 1120–1135.

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* 37(4):1705–1732.

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends Machine Learn.* 5(1):1–122.

Carpentier A, Munos R (2012) Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. Lawrence ND, Girolami M, eds. *Artificial Intelligence and Statistics* (PMLR, Cambridge, MA), 190–198.

Chow SC, Chang M (2008) Adaptive design methods in clinical trials—A review. *Orphanet J. Rare Diseases* 3(1):1–13.

Chu W, Li L, Reyzin L, Schapire R (2011) Contextual bandits with linear payoff functions. Gordon G, Dunson D, Dudík M, eds. *Proc. 14th Internat. Conf. on Artificial Intelligence and Statist.* (PMLR, Cambridge, MA), 208–214.

Cover TM, Thomas JA (2006) *Elements of Information Theory*, 2nd ed. (Wiley, New York).

Ferreira KJ, Simchi-Levi D, Wang H (2018) Online network revenue management using thompson sampling. *Oper. Res.* 66(6):1586–1602.

Filippi S, Cappe O, Garivier A, Szepesvári C (2010) Parametric bandits: The generalized linear case. Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 586–594.

Gao Z, Han Y, Ren Z, Zhou Z (2019) Batched multi-armed bandits problem. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 501–511.

Goldenshluger A, Zeevi A (2013) A linear response bandit problem. *Stochastic Systems* 3(1):230–261.

Han Y, Zhou Z, Zhou Z, Blanchet J, Glynn PW, Ye Y (2020) Sequential batch learning in finite-action linear contextual bandits. Preprint, submitted April 14, https://arxiv.org/abs/2004.06321.

Hastie T, Tibshirani R, Wainwright M (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations* (CRC Press, Boca Raton, FL).

Hopp WJ, Li J, Wang G (2018) Big data and the precision medicine revolution. *Production Oper. Management* 27(9):1647–1664.

Joachims T, Swaminathan A, Rijke MD (2018) Deep learning with logged bandit feedback. Bengio Y, LeCun Y, eds. *Proc. Internat. Conf. on Learn. Representations* (OpenReview, Amherst, MA).

Kallus N, Zhou A (2018) Confounding-robust policy improvement. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Advances in neural information processing systems* 31 (Curran Associates, Inc, Red Hook, NY).

Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, Stewart DJ, et al. (2011) The battle trial: Personalizing therapy for lung cancer. *Cancer Discovery* 1(1):44–53.

Kim GS, Paik MC (2019) Doubly-robust lasso bandit. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 5877–5887.

Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.

Lattimore T, Szepesvári C (2020) *Bandit algorithms* (Cambridge University Press, Cambridge, UK).

Li K, Yang Y, Narisetty NN (2021) Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit. *Electronic J. Statist.* 15(2):5652–5695.

Miao S, Chao X (2019) Fast algorithms for online personalized assortment optimization in a big data regime. Preprint, submitted August 5, https://dx.doi.org/10.2139/ssrn.3432574.

Mintz Y, Aswani A, Kaminsky P, Flowers E, Fukuoka Y (2017) Behavioral analytics for myopic agents. Preprint, submitted February 17, https://arxiv.org/abs/1702.05496.

Mintz Y, Aswani A, Kaminsky P, Flowers E, Fukuoka Y (2020) Nonstationary bandits with habituation and recovery dynamics. *Oper. Res.*

Naik P, Wedel M, Bacon L, Bodapati A, Bradlow E, Kamakura W, Kreulen J, et al. (2008) Challenges and opportunities in high-dimensional choice data analyses. *Marketing Lett.* 19(3–4):201.

Oh Mh, Iyengar G, Zeevi A (2021) Sparsity-agnostic lasso bandit. Meila M, Zhang Tong, eds. *Proc. Internat. Conf. on Machine Learn.* (PMLR, Cambridge, MA), 8271–8280.

Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, Holmes J, et al. (2018) Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Medicine* 16(1):1–15.

Perchet V, Rigollet P, Chassang S, Snowberg E (2016) Batched bandit problems. *Ann. Statist.* 44(2):660–681.

Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D (2015) Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* 3(4):277–287.

Rigollet P (2015) *18. s997: High Dimensional Statistics* (MIT OpenCourseWare, Cambridge, MA).

Rigollet P, Zeevi A (2010) Nonparametric bandits with covariates. Kalai AT, Mohri M, eds. *Conference On Learning Theory* (Omnipress, Norristown, PA), 54–66.

Robbins H (1952) Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc. (New Series)* 58(5):527–535.

Rudelson M, Vershynin R (2015) Small ball probabilities for linear images of high-dimensional distributions. *Internat. Math. Res. Not. IMRN* 2015(19):9594–9617.

Russo D, Van Roy B (2016) An information-theoretic analysis of Thompson sampling. *J. Machine Learn. Res.* 17(1):2442–2471.

Schwartz EM, Bradlow ET, Fader PS (2017) Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Sci.* 36(4):500–522.

Slivkins A (2019) Introduction to multi-armed bandits. *Foundations Trends Machine Learn.* 12(1–2):1–286.

Swaminathan A, Joachims T (2015) Batch learning from logged bandit feedback through counterfactual risk minimization. *J. Machine Learn. Res.* 16(52):1731–1755.

Wainwright MJ (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48 (Cambridge University Press, Cambridge, UK).

Wang X, Wei M, Yao T (2018) Minimax concave penalized multi-armed bandit model with high-dimensional covariates. *Proc. Internat. Conf. on Machine Learn.*, 5200–5208.

Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR (2014) Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* 102(1):151–168.

Zhou M, Fukuoka Y, Mintz Y, Goldberg K, Kaminsky P, Flowers E, Aswani A (2018) Evaluating machine learning–based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial. *JMIR Mhealth Uhealth* 6(1):e28.